

Open AI Governance and Deployment Framework

An operating manual for institutions that buy, deploy, and run AI systems

Version 1.0

1 October 2026

[doi:10.5281/zenodo.22986774](https://doi.org/10.5281/zenodo.22986774)

Published by Yellow Sapphire Technologies Inc. (YS Tech), a New Brunswick corporation with its registered office in Moncton, Canada, and offices in Dubai, Doha, London, and New Delhi.

Published under the Creative Commons Attribution 4.0 International licence (CC BY 4.0). You are free to use, adapt, translate, and redistribute this document, provided you give attribution. Details in the back matter.

This framework is guidance and it is not legal advice. It was not endorsed, approved, or authored by the United Nations. Full disclaimers in the back matter.

Contents

Foreword.....	4
How to use this framework.....	5
Executive summary.....	6
Module 1: Foundations.....	7
1.1 The deployer's seat.....	7
1.2 The five commitments.....	7
1.3 Where this framework sits among the standards.....	8
1.4 What this framework does not cover, and when to get help.....	9
Artifact 1: The one-page orientation.....	10
References for this module.....	11
Module 2: Risk tiering.....	12
2.1 Why tiering comes first.....	12
2.2 The method.....	12
2.3 The four tiers and what each one requires.....	13
2.4 Edge cases and rules.....	14
2.5 Worked classifications.....	15
2.6 Where this module sits in existing standards.....	17
Artifact 2: Risk classification worksheet.....	18
References for this module.....	19
Module 3: Governance roles and human oversight.....	20
3.1 What broken oversight looks like.....	20
3.2 Appoint the four roles.....	20
3.3 Set the four gates.....	21
3.4 Make oversight real.....	22
3.5 The rubber-stamp trap, and the countermeasures.....	23
3.6 Escalation, and the pause rule.....	23
3.7 Where this module sits in existing standards.....	24
Artifact 3a: Governance RACI.....	25
Artifact 3b: Role definition pack.....	26
References for this module.....	27
Module 4: Procurement and vendor accountability.....	28
4.1 Why procurement is the control point.....	28
4.2 What to demand before signing.....	29
4.3 The contract terms that matter.....	31
4.4 Questions that expose a weak vendor.....	34
4.5 Red flags that stop a purchase.....	35
4.6 Scaling this module to the risk tier.....	36
4.7 Where this module sits in existing standards.....	37
Artifact 4a: Vendor due diligence checklist.....	38
Artifact 4b: Contract clause library.....	40
References for this module.....	42
Module 5: Data, security, and sovereignty.....	43
5.1 What you control, and what you hand over.....	43
5.2 Inventory what would leave the building.....	43
5.3 Settle where the data goes, and whose law follows it.....	44
5.4 Security controls that matter for AI systems specifically.....	45
5.5 Send less, and put an end date on everything.....	46

5.6 Special categories.....	46
5.7 Where this module sits in existing standards.....	47
Artifact 5: Pre-deployment data checklist.....	48
References for this module.....	49
Module 6: The deployment lifecycle.....	50
6.1 The life of a system.....	50
6.2 Validation.....	51
6.3 Phased rollout.....	51
6.4 Monitoring.....	52
6.5 Complaints and appeals from affected persons.....	53
6.6 Incident response.....	53
6.7 Vendor updates.....	54
6.8 Periodic revalidation.....	55
6.9 Decommissioning.....	55
6.10 Where this module sits in existing standards.....	56
Artifact 6a: Stage-by-stage lifecycle checklist.....	57
Artifact 6b: Monitoring plan template.....	58
Artifact 6c: Appeal notice template.....	59
References for this module.....	59
Module 7: The implementation toolkit.....	60
7.0 The thirty-day path.....	60
7.0.1 What the completed toolkit evidences.....	61
7.1 The one-page orientation (from Module 1).....	62
7.2 Risk classification worksheet (from Module 2).....	63
7.3 Vendor due diligence checklist (from Module 4).....	65
7.4 Contract clause library (from Module 4).....	67
7.5a Governance RACI (from Module 3).....	70
7.5b Role definition pack (from Module 3).....	71
7.6 Pre-deployment readiness pack (gates G2 and G3).....	72
7.6a Pre-deployment data checklist (from Module 5, attachment A to the readiness pack).....	73
7.7 Stage-by-stage lifecycle checklist (from Module 6).....	75
7.8 Monitoring plan template (from Module 6).....	76
7.9 Appeal notice template (from Module 6).....	76
7.10 Incident response template.....	77
7.11 The system register.....	78
References for this module.....	78
Annex A: Full mapping to external standards.....	79
Annex B: Glossary.....	81
Annex C: References.....	84
The limits of this framework, restated.....	85
About Yellow Sapphire Technologies.....	85
How to cite this framework.....	85
Licence.....	85
Disclaimers.....	85
Feedback and versioning.....	86

Foreword

Most of what has been written about governing artificial intelligence is addressed to the organisations that build AI models. Almost none of it till now is addressed to the hospital, the ministry, the utility, the bank, or the school system that buys an AI system, installs it, and runs it inside decisions that affect real people. Yet that is where most of the world will meet AI: through institutions that deploy it, staffed by capable people who likely have no budget for an AI safety team, no model evaluation function, often no in-house counsel, and no time.

This framework is written for those people, from their seat. It is an operating manual. The principles are set out in international instruments. What an overstretched technology or operations lead needs is the translation layer; the questions to put in a tender, the clauses that make a vendor's claims enforceable, the names on the sign-off, the numbers on the monitoring plan, and the route by which an affected person can say the system got it wrong.

In 2026, Yellow Sapphire Technologies was invited to contribute to the United Nations Global Dialogue on AI Governance, and we registered this framework as a public commitment after the Dialogue. This document is the delivery of that commitment. It is published free of charge, under an open licence, for any government, enterprise, or civil society organisation that can use it.

Each recommendation is written to be specific enough to act on, and each module ends with an artifact that an institution can adapt and put to work. Additionally, we want to acknowledge we don't know what we don't know. The framework says what it cannot do, where national law governs, and when an institution needs specialist help. It draws on documented cases and established standards, and it maps onto the recognised international instruments.

Use it. Adapt it to your institution, translate it into your language, localise it to your law, and improve it. The licence permits all of this. Version 1.0 is just a beginning; I will personally oversee the feedback channel so we can improve this framework further. Its purpose is to help an institution refuse a bad contract, catch a silent failure, and hear the person its system wronged.

Anuuj Chauhan

Co-Founder and CEO, Yellow Sapphire Technologies
Delhi, 1 October 2026

How to use this framework

Who this is for. A technology or operations lead inside an institution adopting AI: a public hospital, a ministry, a municipal utility, a regional bank, a tax authority, a school system, a mid-sized enterprise. It assumes no AI expertise, no compliance department, and no budget for consultants. It covers systems you buy and run, both predictive (systems that score, rank, classify, or forecast) and generative (systems that produce text, images, or code). Systems authorised to act on their own (agentic systems) are covered through Module 2's autonomy rules, which treat an adverse action taken without prior human review as a matter for escalation.

The structure. Seven modules. Module 1 is the fifteen-minute on-ramp: what the framework commits to and how it maps to recognised standards. Module 2 classifies each AI use case into a risk tier in an afternoon; the tier doses everything after it. Module 3 names the four governance roles and makes human oversight real. Module 4, the largest, is procurement: what to demand, ask, and sign before a vendor contract. Module 5 settles data, security, and sovereignty before anything leaves the building. Module 6 runs the deployment lifecycle: validation on your own data, phased rollout, monitoring, appeals from affected people, incidents, vendor updates, and switch-off. Module 7 consolidates every artifact and is where the work starts.

Reading paths.

- **A purchase is live this month:** read Module 2, then Module 4. Use the risk worksheet, the due diligence checklist, and the clause library as you go.
- **Systems are already running with no governance around them:** read Module 2, then Module 3, then Module 6 sections 6.4 and 6.5. Classify, name owners, start monitoring and the appeal channel.
- **Building from zero with time to plan:** read the modules in order, then run the thirty-day path.

The thirty-day path (Module 7.0). Week 1: find every AI use in the institution, including free tools, and give each a row in the system register. Week 2: classify every row with the risk worksheet, each classification confirmed by a second person. Week 3: take the one-page orientation to leadership and get owners named for every classified system, with all four roles for the high-stakes ones. Week 4: put the first controls on the riskiest live system: a monitoring plan, the published appeal notice, and a contract check against the clause library. Day 30: a one-page status to leadership. The full path, with its outputs and its limits, is section 7.0.

Conventions. Technical terms are defined where first used and collected in the glossary at the back. “Check locally” marks every point where national law governs and tells you what to ask. Bracketed items in artifacts are entries you must fill in. Operational numbers in this framework (hours, durations, sample sizes, thresholds, clocks) are **framework defaults** unless a law or cited source is named beside them: recommended starting values, chosen for workability, that your institution adopts, adjusts, and records. They are design choices of this framework rather than empirical findings or legal requirements. “Monday morning” means the start of your working week, whichever day that is. Every artifact, once adopted, carries a document control block: version, prepared by, last updated, next review. All artifacts are also published as separate editable files at ystech.io/ai-governance. The points where paid legal or specialist review is required are named in the text, and there are few of them.

Executive summary

The argument. AI governance is being written for the organisations that build AI. It will be lived by the organisations that deploy it. Nearly all published guidance addresses the developers of models, with their evaluation teams and compliance departments. Almost none of it addresses the institution that buys a system, installs it, and runs it inside decisions about care, money, benefits, education, or essential services. That institution typically has a procurement office, an IT team at capacity, a vendor's sales deck, and a deadline. The gap is visible in the instruments themselves: most of the EU AI Act's obligations attach to providers of systems. This framework fills that gap: an operational manual for adopting and running AI responsibly, written for the deployer's seat.

Why it matters. A sepsis prediction tool deployed at hundreds of hospitals on claimed accuracy of 0.76 to 0.83 measured 0.63 when one buyer tested it on its own patients, missing two thirds of sepsis cases; a question in the tender about the population behind those figures, and a test on the buyer's own records before go-live, would have exposed the gap. A widely used care management algorithm systematically underrated the illness of Black patients because it predicted healthcare spending rather than need; correcting it would have raised the share of Black patients identified for extra help from 17.7 to 46.5 percent. And a national tax administration ran a fraud scoring model over benefit claims with governance suited to an internal tool; roughly 26,000 parents were wrongly accused, and the government resigned over it.

What the framework commits to. Five operating rules, each turned into checklists, clauses, and signatures: accountability names a person; evidence comes before trust; oversight carries real authority; watching continues after go-live; and you can always leave.

How it works. Every AI use case is classified into a risk tier in an afternoon, and the tier sets the governance dose. Four named roles, held by existing staff alongside their jobs, carry accountability through four sign-off gates. Procurement is treated as the point of maximum power: the framework specifies the documents to demand, the sixteen questions that expose a weak vendor, the red flags that stop a purchase, and a plain-language library of fourteen contract clauses. Data questions are settled before anything leaves the building. Every high-stakes system is tested on the institution's own population before go-live, rolled out in phases, monitored against thresholds written in advance, and revalidated yearly. Affected people get a published route to challenge decisions, with deadlines. Vendor model updates are re-tested against a frozen test set before acceptance. Every system has an exit. A consolidated toolkit of eleven artifacts, also available as editable files, makes all of it liftable, and a thirty-day path takes one person from an unclassified estate to a signed register with named owners.

What it asks of leadership. Its first ask is one line of authority: a written directive adopting the framework and mandating its thirty-day path, because everything below flows from that mandate. Then four things: appoint the four roles (about 14 hours per month across them for one high-stakes system, a framework default); require the risk worksheet for every AI use, including free tools; require the procurement gate before any AI contract is signed; and receive a one-page monitoring summary monthly. What it prevents: buying systems that do not work on your population, contracts you cannot exit, silent degradation in service, and standing unprepared before a regulator, an auditor, or an affected person.

What it is not. It restates no principles and competes with no standard. It maps visibly onto the EU AI Act, the NIST AI Risk Management Framework, ISO/IEC 42001, the OECD AI Principles, and the UNESCO Recommendation on the Ethics of AI, so work done under it is evidence of movement towards recognised standards, and where any law applies, the law prevails. It states what it cannot do and names the few points where an institution must buy specialist help.

Provenance and licence. Written and published by Yellow Sapphire Technologies Inc. as the delivery of a commitment registered through the Partnerships Hub of the United Nations Global Dialogue on AI Governance; it was not endorsed, approved, or authored by the United Nations. Published free of charge under CC BY 4.0: any institution may use, adapt, translate, and localise it, with attribution.

Module 1: Foundations

You have been handed responsibility for AI systems, along with governance material written for the companies that build them. This module gives you the map for your seat: what this framework covers, the five commitments it runs on, how it maps to the recognised standards, and a statement of what it will not do for you.

It is the shortest module in the framework and it is the on-ramp, a quick fifteen-minute read. If a purchase is live right now, go to Module 2 to set the risk tier, then Module 4 for the procurement work, and come back here later.

1.1 The deployer's seat

A vendor has demonstrated a system to your leadership: a tool that scores loan applications, or flags patients for follow-up, or drafts responses to citizens. Leadership is interested. Procurement has a deadline. Someone senior has asked you to “make sure the governance side is covered,” and the material you find when you look is written for a different reader: it addresses the companies that train models, with their evaluation teams, safety researchers, and compliance departments. You have an IT team at capacity and a tender to issue by the end of the month.

This framework is written for your seat.

A company builds a system that predicts which hospital patients need extra care. The builder's obligations run to the system itself: assemble the training data, test the model, document its accuracy, write the instructions. A hospital then buys that system and turns it on. From that moment a second set of obligations exists, and they belong to the hospital alone, because only the hospital can perform them:

- Deciding whether this use, on these patients, is high stakes (Module 2).
- Naming the person accountable when the system is wrong, and the person with authority to switch it off (Module 3).
- Extracting the evidence, the contract terms, and the exit rights from the vendor before signing (Module 4).
- Settling what patient data leaves the building, to which country, under whose control (Module 5).
- Testing the system on the hospital's own patients before relying on it, watching it in service, and hearing the patient who says the decision was wrong (Module 6).

Nearly all published governance material addresses the builder's list. The pattern is visible in the instruments themselves: most of the EU AI Act's obligations attach to providers of systems. This framework addresses the second list. It is an operating manual for institutions that buy, deploy, and run AI systems: hospitals, ministries, utilities, banks, schools, municipalities, and enterprises. It assumes you have no AI specialists.

Scope. The framework covers acquired systems of both kinds: predictive systems, which score, rank, classify, or forecast; and generative systems, which produce text, images, or code, including chat assistants and drafting tools. It governs the buying and the running. Systems authorised to act on their own (agentic systems: executing transactions, sending messages, changing records without a human touching each action) are inside scope and are handled through Module 2's autonomy rules, which treat an adverse action taken without prior human review as a matter for escalation. If your institution builds or fine-tunes its own models, you hold both lists of obligations at once: this framework still governs your deployment, and for the building you should work directly from the NIST AI Risk Management Framework and ISO/IEC 42001, both introduced in section 1.3.

1.2 The five commitments

Everything in this framework operationalises five commitments. Each one lives in a specific module where it becomes checklists, clauses, and named signatures.

1. Accountability names a person. Every AI system in your institution has one named individual who is accountable for its behaviour in service and who has the authority to stop it. Module 3 defines the role, and the system register in Module 7 records the name.

2. Evidence comes before trust. No claim about a system is accepted without a document behind it, and no high-stakes system goes live without being tested on your institution's own data. Module 4 extracts the evidence from the vendor; Module 6 runs the test.

3. Oversight carries real authority. A human overseeing an AI system has the power to override it, the training to know when to, and the time to look. Module 3 builds this.

4. Watching continues after go-live. Module 6 sets what to measure in service, how often, who looks, and the threshold at which someone must act, and it gives affected people a named channel to challenge a decision, with a deadline for your answer.

5. You can always leave. Every system has an exit: contractual rights to your data and a plan for switching off. Module 4 writes the exit into the contract; Module 6 executes it.

1.3 Where this framework sits among the standards

This framework restates no principles and competes with no standard. Five established instruments matter to a deployer, and this framework is the translation layer between them and your Monday morning:

- **The EU AI Act**, adopted in 2024, the first comprehensive AI law, whose obligations phase in over several years. Its risk-based logic (stricter duties for higher-stakes uses) is the logic of Module 2, and it places obligations on deployers as well as providers.
- **The NIST AI Risk Management Framework**, published by the United States National Institute of Standards and Technology in 2023. A voluntary framework organised around four functions: Govern, Map, Measure, and Manage. Its structure underlies Modules 2, 3, and 6.
- **ISO/IEC 42001**, published in December 2023 by the International Organization for Standardization and the International Electrotechnical Commission (jointly, ISO/IEC), the first international management system standard for AI. It describes what an organisation-level AI management system contains. This framework is a lightweight on-ramp to it.
- **The OECD AI Principles** (the Organisation for Economic Co-operation and Development), adopted in 2019 and updated in 2024, the first intergovernmental standard on AI.
- **The UNESCO Recommendation on the Ethics of Artificial Intelligence** (the United Nations Educational, Scientific and Cultural Organization), adopted by UNESCO's member states in November 2021.

The last two are values instruments.

The mapping, at concept level (precise article and clause references are collected and verified in the annex):

Module in this framework	EU AI Act	NIST AI RMF	ISO/IEC 42001	OECD Principles and UNESCO Recommendation
2: Risk tiering	The Act's risk-based structure: obligations scale with the stakes of the use	Map: establishing context and classifying risk	Risk assessment expectations within the management system	Governance effort scaled to the risk of harm to people
3: Roles and human oversight	The Act's human oversight requirements for high-risk systems, and deployer duties to assign competent oversight	Govern: accountability structures, roles, and culture	Leadership, roles, and responsibilities clauses	Accountability; human determination
4: Procurement and vendors	Provider documentation and instructions for use, which a deployer must obtain and follow	Govern and Map: third-party risk	Supplier and third-party controls	Transparency and accountability, made contractual

Module in this framework	EU AI Act	NIST AI RMF	ISO/IEC 42001	OECD Principles and UNESCO Recommendation
5: Data, security, sovereignty	The Act's data governance concerns; national data protection law sits alongside and prevails	Measure and Manage: data and security risk	Data and asset management expectations	Privacy; safety and security
6: Deployment lifecycle	Deployer duties to monitor operation and report serious incidents	Measure and Manage: test, monitor, respond	Operation, performance evaluation, and improvement clauses	Safety; accountability through the full lifecycle
7: Toolkit	The working evidence that the duties above were performed	The four functions as filled-in documents	Records the management system requires	Documented evidence that the values were applied to each system, before and after go-live

Where a national law or regulation applies to you, it prevails over this framework, and Module 5 tells you which questions to take to local advice.

1.4 What this framework does not cover, and when to get help

1. **It is not legal advice.** It tells you what to demand and what to check; a lawyer qualified in your jurisdiction tells you what the words mean in your country. The framework marks the points where that review is required (contract signature in Module 4, data transfer in Module 5, incident notification in Module 6).
2. **It does not evaluate model internals.** The framework's method is external: evidence, testing on your own data, monitoring in service, and enforceable contracts.
3. **It does not certify compliance** with any law or standard, including the five instruments above. The mapping shows direction of travel. An auditor or regulator confirms arrival.
4. **It does not replace domain expertise.** Whether a clinical tool is medically sound, a credit model is financially sound, or a grid tool is operationally sound remains a judgement for your clinicians, your risk officers, your engineers. The framework gets them the evidence and the authority to make that judgement.
5. **It does not govern the building of models.** Builders and fine-tuners: see the scope note in section 1.1.
6. **It cannot supply capacity you do not have.** It is designed for a stretched team, and Module 2 exists so that effort goes where the stakes are. It still costs hours: roughly 14 hours per month across the four roles for a single high-stakes system in steady operation (a framework default, stated per role in Module 3 for you to adjust and record).

Get specialist help, and budget for it, at these specific points: legal review before any Tier 2 signature; local data protection advice before personal data crosses a border; security testing for any system exposed to the public; independent adversarial testing for systems where successful manipulation could affect safety, security screening, or access to essential services; and domain review (clinical, financial, engineering) before any Tier 2 system passes its go-live gate.

Where to go next. A live purchase: Module 2, then Module 4. Tools staff adopted without procurement, including free tools: Module 2 classifies them and Module 7's first week finds them. Systems already running with no governance around them: Module 2 to tier them, then Module 3 to name the roles, then Module 6 to start monitoring. Building from zero with time to plan: read the modules in order, then run the thirty-day path (Module 7.0).

Artifact 1: The one-page orientation

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0.

Purpose: a single page for a director, board member, minister, or executive whose institution is adopting this framework. It is designed to be sent ahead of the meeting where you ask for the four appointments at the bottom.

AI governance for institutions that deploy: one page

The situation. Our institution buys and runs AI systems; it does not build them. Almost all AI governance guidance is written for builders. This framework is an operating manual for our seat: what to demand from vendors, who is accountable internally, what to check before a system goes live, what to watch afterwards, and how to switch off.

The framework we are adopting. The Open AI Governance and Deployment Framework (Version 1.0), a public document published under an open licence. It maps onto the EU AI Act, the NIST AI Risk Management Framework, ISO/IEC 42001, the OECD AI Principles, and the UNESCO Recommendation on the Ethics of AI, so work done under it is evidence of movement towards recognised standards, though the framework certifies compliance with none of them. It is not legal advice, and it prescribes where legal and specialist review is required.

The five commitments it runs on.

1. Accountability names a person: every system has one named owner with the authority to stop it.
2. Evidence comes before trust: no vendor claim without a document; no high-stakes go-live without a test on our own data.
3. Oversight carries real authority: the humans reviewing AI outputs can override them, and have the training and time to do it.
4. Watching continues after go-live: defined metrics, a defined reviewer, defined thresholds for action, and a channel for affected people to challenge a decision.
5. We can always leave: every contract includes exit and data return; every system has a switch-off plan.

What this asks of leadership.

1. Appoint four roles (existing staff hold them alongside their jobs, with the monthly time cost stated per role): a System Owner per AI system, an Operational Lead, a Data Steward, and an Independent Reviewer for go-live sign-offs.
2. Require the risk classification worksheet for every AI use case, existing and proposed, including free tools.
3. Require the procurement gate: no AI contract is signed until the vendor due diligence checklist passes for the system's risk tier.
4. Receive the monitoring summary: one page per month on every high-stakes system, from the System Owner.

What it costs. Staff hours, stated per role in the framework, and paid specialist help at the points named in Module 1.4: legal review, data protection advice, security testing, adversarial testing, and domain review. What it prevents: buying systems that do not work on our population, contracts we cannot exit, silent degradation in service, and standing unprepared in front of a regulator, an auditor, or an affected person.

Directive and designations. By signing, leadership adopts this framework for the institution and mandates the thirty-day path (Module 7), including the instruction to departments to answer its first-week question.

Role	Name	Date
System Owner for [system]		
Operational Lead		
Data Steward		
Independent Reviewer arrangement		

Authorised by (name, title, signature): _____ Date: _____

Prepared by: _____ For: _____ Date: _____

References for this module

- OECD, Recommendation of the Council on Artificial Intelligence (the OECD AI Principles), adopted May 2019, updated 2024.
- UNESCO, Recommendation on the Ethics of Artificial Intelligence, adopted November 2021.
- ISO/IEC 42001:2023, Information technology, Artificial intelligence, Management system. Published December 2023.
- NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), published January 2023.
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (the EU Artificial Intelligence Act). Entered into force 1 August 2024, with obligations applying in phases from 2025. Amended by Regulation (EU) 2026/1744 of 8 July 2026 (Digital Omnibus on AI), in force 27 July 2026.

Module 2: Risk tiering

Your institution has more AI use cases than governance capacity, and no defensible way to decide which ones get the full treatment. This module gives you a classification method one person can run in an afternoon. It produces a signed record of each system's risk tier, and the tier tells you how much governance the system must carry.

Read sections 2.1 to 2.4 (about twenty minutes), walk through the three worked examples in 2.5, then classify your own systems with the worksheet (Artifact 2). Classify every AI use in the institution, including the free tools staff have adopted on their own, and including systems already live.

2.1 Why tiering comes first

Between 2005 and 2019, the tax administration of the Netherlands wrongly accused roughly 26,000 parents of making fraudulent childcare benefit claims and ordered them to repay benefits in full, in many cases tens of thousands of euros. Families were driven into debt and hardship. Part of the machinery was a self-learning risk classification model: it scored benefit claims for fraud risk, and officials concentrated their scrutiny on the claims the model ranked highest. The Dutch Data Protection Authority later found that the system had used nationality as a risk factor and that the processing was discriminatory and unlawful, and in 2021 fined the tax administration 2.75 million euros. A parliamentary inquiry concluded that fundamental principles of the rule of law had been violated. On 15 January 2021, the Dutch cabinet resigned over the affair. (Sources in the references at the end of this module.)

A model that scored families for fraud, where a high score could trigger investigation, repayment demands, and the loss of benefits, was run with the governance you might give an internal reporting tool. The institution lacked a step where someone wrote down, before the system ran at scale: this system makes adverse calls about people's money and legal standing, so it belongs in our highest risk class, and here is the oversight, the input scrutiny, and the appeal channel that class requires. Everything this framework prescribes downstream (the roles in Module 3, the vendor demands in Module 4, the monitoring and the appeal channel in Module 6) is dosed by that classification.

Your capacity is finite. If every AI use requires the full framework, your institution will quietly stop following the framework, and the decision about what gets governed will be made by fatigue.

Classify the use, and classify it by the riskiest realistic use. The same product can sit in different tiers in different uses: a general-purpose chat assistant is one thing drafting internal minutes and a different thing answering benefit claimants. The tier belongs to the use case. And the tier is set by the riskiest use that will realistically occur, which is a question about your institution's behaviour rather than the vendor's intentions. If the "decision support" tool's suggestions will, on a busy Friday, be accepted unread, classify the Friday version.

2.2 The method

The method is designed to take one system and one afternoon (a framework default; your first few will run longer). You will check one list and answer the questions below, in order, and the answers place the use case in one of four tiers. A second person then confirms the classification; Step 5 says who.

Step 1. Check the escalation list. A small number of uses should never be classified and deployed on your own authority, because they are legally restricted or outright prohibited in a growing number of jurisdictions, and because their failure modes exceed what this framework's controls can contain. If the use case does any of the following, stop here and mark it **Tier 1: escalate**:

- Scores or ranks people's general trustworthiness or social behaviour across contexts (social scoring).
- Identifies people biometrically in public spaces, or categorises people by biometric data.
- Infers emotions, or infers protected characteristics such as ethnicity, religion, or sexual orientation, about employees, students, or the public.

- Makes adverse decisions about people fully automatically, with no human review of the individual decision and no route of appeal.
- Is known to be prohibited or restricted by a law that applies to you; if you do not know, do not decide: mark the status unknown and route the worksheet to legal review. **Check locally:** the EU AI Act prohibits several practices of this kind and heavily regulates others, and other jurisdictions have their own lists. Your escalation list should grow to match your law; it should never shrink below the list above.

A Tier 1 use proceeds only after review by qualified legal counsel and a relevant specialist, on a written record, with leadership's explicit sign-off, and this framework alone is insufficient for it.

Step 2. The decision question. *Does the system inform or make decisions that can be adverse to an identifiable person?* Adverse means the person can lose something that matters: access to care, money or credit, benefits, housing, education, employment, essential services, legal standing, or liberty. "Inform" counts the same as "make": a score that a human reads before deciding still shapes the decision.

If yes: the use case is **Tier 2: high stakes**. Go to the modifiers in Step 4.

Step 3. The dependence question. *Do the system's outputs steer operations that people depend on, without deciding about any identifiable person?* Demand forecasts, maintenance prioritisation, network optimisation, inventory and staffing projections: nobody is scored, but the lights, the water, the supplies, and the rosters follow the output.

If yes: **Tier 3: operational**. If no, meaning the system's outputs stay inside a workflow where a wrong output costs rework rather than harm (drafting aids, internal search, summarisation, translation for internal use, code assistance with human review): **Tier 4: low stakes**. Go to Step 4 either way.

Step 4. Two modifiers. These move a case up a tier. Nothing moves a case down.

- **The autonomy modifier.** Will individual outputs take effect without a human reviewing them first? At Tier 3 or Tier 4, this moves the case up one tier. At Tier 2, an adverse decision about a person with no individual human review belongs on the escalation list in Step 1; treat it as Tier 1. **Check locally:** many data protection laws restrict fully automated decisions with significant effects on people and grant rights to human review.
- **The detectability modifier.** If the system produced a wrong output, would anyone notice within days? A forecast is checked against reality every morning. A patient the triage tool failed to flag is invisible because nobody was told to look. If wrong outputs would likely go unnoticed, move a Tier 3 or Tier 4 case up one tier. At Tier 2, low detectability changes the controls rather than the tier: it mandates the strict path in Module 6 (shadow mode before go-live, and monitoring designed to count the failures that produce no alert, both defined there).

Step 5. Record and sign. Write the answers on the worksheet (Artifact 2), with one sentence of reasoning per step. The person who classified signs. A second person confirms and signs: for Tier 2, the System Owner named under Module 3; below Tier 2, a peer outside the proposing team. File the worksheet in the system register (Module 7) and set the review date from the table below.

2.3 The four tiers and what each one requires

- **Tier 1, escalate.** A social scoring proposal; workplace emotion monitoring; fully automated benefit denials. Requires legal and specialist review before anything else happens.
- **Tier 2, high stakes.** A patient deterioration alert in a hospital; a fraud risk flag in a benefits ministry; loan pre-screening in a bank; an exam grading tool in a school system; a chatbot answering the public's questions about their own entitlements.
- **Tier 3, operational.** An electricity load forecast in a utility; predictive maintenance on water pumps; demand forecasting for medical supplies; staffing projections.
- **Tier 4, low stakes.** A meeting minutes summariser; internal document search; a drafting assistant whose output a named human edits and owns; code completion with human review.

Why four tiers: three collapses the operational middle, so forecasting tools would share a tier either with drafting aids or with adverse decisions about people, and both mistakes are costly in opposite directions. Five or more exceeds what one person can classify in an afternoon and what an institution can realistically staff differently. Four is the smallest number that separates the uses this framework must treat differently. It is a design choice of this framework, and where a law that applies to you defines its own categories, the law's categories govern.

The governance load, by tier. Each cell points to the module that defines the requirement.

Requirement	Tier 2	Tier 3	Tier 4
Classification worksheet signed and confirmed (this module)	Yes, confirmed by the System Owner	Yes, peer confirmed	Yes, peer confirmed
Named roles (Module 3)	All four roles	System Owner, Operational Lead, and Data Steward	System Owner (can hold a portfolio of Tier 4 tools) and Operational Lead
Procurement path (Module 4)	Full module, checklist at Tier 2 threshold	Documentation core plus six clauses (Module 4.6)	Short path: three questions, three clauses (Module 4.6)
Data checklist before deployment (Module 5)	Full checklist, signed by the Data Steward	Full checklist	Data questions from the short path
Pre-deployment validation on your own data (Module 6)	Required; shadow mode first where detectability is low	Required, lightweight	Recommended for tools in wide use
Monitoring in service (Module 6)	Monthly review as default; weekly where detectability is low; thresholds written before go-live	Monthly against actuals	Annual sanity check
Appeal channel (Module 6)	Required, published to affected people	Required only if outputs later touch people (which re-tiers the case)	No
Incident response plan (Module 6)	Required before go-live	Required, may be the institution's standard IT process plus the Module 6 additions	Institution's standard IT process
Exit plan and contract exit rights (Modules 4 and 6)	Required before signature	Required before signature	Export and deletion confirmed before rollout
Re-classification review	Every 12 months and on any trigger in 2.4	Every 12 months and on any trigger	Every 24 months and on any trigger

For a single Tier 2 system in steady operation, the recurring load in this table is measured in hours per month, and Module 3 states the hours per role. The one-time load (validation, go-live) runs from contract signature to full go-live in typically three to six months of elapsed time for a Tier 2 system (a framework default), with the stage durations set in Module 6. If leadership wants the system live sooner, the table is the list of what they are choosing to skip, in writing.

2.4 Edge cases and rules

1. General-purpose tools are classified per use, and the uses multiply. A chat assistant procured for drafting (Tier 4) will be used, within weeks, by someone answering a citizen's query or summarising a case file. Handle this by pairing every general-purpose deployment with a one-paragraph acceptable-use note derived from its tier: written in the language of tasks ("use it to draft and summarise internal documents; do not paste personal data into it; do not use its output as the answer given to a member of the public without a named person checking it"). The note is shown to every user when their account is created and again at first login, and the Operational Lead keeps the current version. When a new use appears that the note excludes, that use gets its own worksheet. Finding the tools nobody procured is Module 7's first week: the software asset inventory, the identity provider's

list of connected applications, accounts payable records, and a direct question to every department head, repeated once. The rule that makes discovery matter is enforceable because it names the data rather than the tool: tools with no data commitments get no institutional data (Module 4.6).

2. Aggregation moves tiers. One AI-drafted reply is Tier 4. The same tool wired to answer all correspondence, with humans sampling a fraction, has become the institution's voice and decision channel, and belongs in Tier 2. Volume and integration change the classification even when the product is unchanged.

3. Advice is a decision when people act on it. A public-facing chatbot that "only provides information" about benefits, health services, or legal processes is making Tier 2 outputs the moment people arrange their affairs around its answers. Classify it by what the reader does with the answer.

4. The tier is set by realistic use. Expect scope creep, plan for it, and catch it with triggers. Re-classify, using a fresh worksheet, when any of these occurs:

- The system starts informing a new kind of decision, or its outputs reach a new group (especially the public).
- A human review step is removed, reduced, or has quietly become nominal (override rates near zero over a sustained period; Module 3.5).
- The vendor adds capabilities, or a model update changes behaviour (Module 6).
- Usage volume grows past what the original worksheet assumed.
- An incident or a pattern of complaints arrives (Module 6).

5. Inputs can end the analysis early. If the use case's inputs include protected characteristics (nationality, ethnicity, religion, and the categories your law protects) or close proxies for them, in a system that touches decisions about people, stop and take it to legal review before classifying further. The Dutch case in 2.1 is the reason this rule exists, and "the vendor configured the inputs" transfers none of the accountability. If the vendor will not disclose the input list, rule 5 cannot be cleared: the classification stops at escalation until it is, and Module 4.5 treats that refusal as a red flag.

6. Pipelines are classified by their end decision. Where several components feed one another (text extraction into summarisation into scoring into human review), classify the pipeline as one use case, by the decision at its end, and validate it end to end (Module 6). A benign component does not lower the tier of a consequential pipeline.

7. Nobody classifies their own project unchecked. The person proposing a system wants it to be Tier 3. The confirming signature in Step 5 exists for this, which is why the confirmer must be someone outside the proposing team for anything below Tier 2.

2.5 Worked classifications

Each example walks the worksheet in order.

Worked example A: patient deterioration alerts in a hospital

A vendor offers a tool that runs on the hospital's electronic records and alerts ward staff when a patient's condition is likely to deteriorate.

- **Step 1, escalation list:** no item applies.
- **Step 2, decision question:** yes. The alert directs clinical attention. A patient the tool should have flagged and did not can receive care later than they otherwise would. Access to care is on the adverse list. **Tier 2.**
- **Step 4, autonomy:** outputs are advisory; clinicians decide. No move.
- **Step 4, detectability:** low, and the worksheet should say why in one sentence: an alert that fires is visible, and a patient who was never flagged is invisible, because the tool's silence looks identical whether it is right or wrong. Tier stays 2; the strict Module 6 path is mandated: shadow mode before go-live (the tool runs, staff decide as they always did, and its alerts are compared against what happened), and in-service monitoring built around missed cases, with the subgroup breakouts from Module 4.2.2.

- **Step 5:** classified by the head of nursing informatics, confirmed and signed by the System Owner, filed, review in 12 months.
- **What the tier buys:** all four Module 3 roles; the full Module 4 procurement, where question 1 of the vendor script (validated on what population?) is the question the sepsis case in Module 4.1 answers; the Module 5 data checklist before any patient record leaves the building; and domain review by clinicians before the go-live gate, per Module 1.4.

Worked example B: a fraud risk flag in a benefits ministry

The ministry considers a system that scores benefit claims for fraud risk so investigators can prioritise.

- **Step 1, escalation list:** run rule 5 of section 2.4 first, because this category has the history: demand the input list from the vendor (Module 4.2.1) and check it for protected characteristics and proxies. If nationality, ethnicity, or stand-ins for them appear, stop and go to legal review. Assume for this example the input list comes back clean.
- **Step 2, decision question:** yes. A flag triggers investigation, and can lead to suspended payments and repayment demands. Money, benefits, and legal standing are on the adverse list. **Tier 2.**
- **Step 4, autonomy:** the ministry's design says an official reviews every flag. Write down the arithmetic on the worksheet: 400 flags a week across 9 investigators is roughly 9 reviews per investigator per day, which is plausible review; 4,000 flags a week is rubber-stamping by design, and under this modifier that is an unreviewed adverse decision, which is Tier 1 escalation territory.
- **Step 4, detectability:** low for the person flagged, who experiences the investigation and cannot see the system. This mandates the strict Module 6 path. The appeal channel is a detection instrument for this system. The claimant's route to challenge, and the ministry's deadline to answer, get designed before go-live.
- **Step 5:** classified by the head of the investigations unit, confirmed by the System Owner, filed, review in 12 months, with override rates (how often investigators disagree with the flag) on the monthly monitoring summary from day one.
- **What the tier buys:** the same load as example A, plus the input scrutiny above. The Dutch case from 2.1 sat in this category. The classification worksheet is short; what it would have recorded, in writing, before scale, is that this system makes adverse calls about families' money, that its input list must survive legal review, and that the people it flags need a way to be heard.

Worked example C: an electricity load forecast in a utility

The utility considers a model that forecasts demand 24 to 72 hours ahead to plan generation and grid operations.

- **Step 1, escalation list:** no item applies.
- **Step 2, decision question:** no. The forecast decides about the grid, and no output attaches to an identifiable person.
- **Step 3, dependence question:** yes. Operations that people depend on follow the output. **Tier 3.**
- **Step 4, autonomy:** planners review the forecast before committing resources. No move. (If the forecast fed an automatic dispatch system with no human step, the modifier would move this to Tier 2, and the worksheet would say so.)
- **Step 4, detectability:** high, and the worksheet says why: yesterday's forecast is checked against yesterday's actual demand every morning, so errors surface in hours. No move.
- **Step 5:** classified by the planning engineer, confirmed by a peer department head, filed, review in 12 months.
- **What the tier buys:** a System Owner, an Operational Lead, and a Data Steward; the Module 4 Tier 3 procurement path (documentation core plus six clauses); the Module 5 checklist (which is quick here, since consumption data at feeder level is the likely payload rather than personal data); monthly monitoring against actuals.

- **The re-tier trigger, named in advance on the worksheet:** if this forecast ever starts driving decisions attached to households or businesses (remote disconnection, load shedding priorities by area, dynamic pricing), it crosses into Step 2 territory and gets a new worksheet as Tier 2.

2.6 Where this module sits in existing standards

This module	Where it maps
The four-tier structure	The EU AI Act's risk-based architecture: prohibited practices, high-risk systems with heavy duties, limited-risk transparency cases, and minimal-risk uses. The Act's categories are legal classifications with legal consequences; this module's tiers are an operating translation for institutions the Act calls deployers, and where the Act applies to you, its classification governs and a lawyer confirms it.
Step 1's escalation list	The Act's prohibited and tightly regulated practices, at concept level; precise references in the annex.
Steps 2 to 4	The NIST AI RMF Map function: establishing context, understanding impact, and categorising risk before measuring or managing it.
Step 5's record, confirmation, and review cycle	ISO/IEC 42001's expectations that risk assessment is documented, repeatable, and periodically refreshed within a management system.
Effort scaling with consequence	The proportionality running through the OECD AI Principles and the UNESCO Recommendation: stewardship appropriate to context and impact.

Artifact 2: Risk classification worksheet

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0. One worksheet per use case, including free tools and systems already live. It is designed to be completed in under an hour once the system description (Module 4, section 4.2.1) is in hand. It is an operating tool and it is not legal advice; where a law classifies your system, the law's classification governs.

Use case being classified

System and vendor	
The use case, in one sentence (what decision or output, for whom)	
Users of the output	
People affected by the output	
Expected volume (outputs per week)	
Model or product version reviewed	
Deployment location(s) and sites	

Step 1: Escalation list. Tick any that apply.

- ☐ Scores or ranks people's general trustworthiness or social behaviour across contexts
- ☐ Biometric identification in public spaces, or biometric categorisation of people
- ☐ Infers emotions or protected characteristics about employees, students, or the public
- ☐ Adverse decisions about people, fully automated, with no individual human review and no appeal
- ☐ Inputs include protected characteristics or close proxies, in a system touching decisions about people (rule 5)
- ☐ Known to be prohibited or restricted by a law that applies to us (state which), or legal status unknown (if unknown, do not decide: route this worksheet to legal review before classifying further): _____

Any tick: **STOP. Tier 1. Route to legal counsel and a relevant specialist with this worksheet attached. Leadership sign-off required to proceed on any path.**

Step 2: Decision question. Does the system inform or make decisions that can be adverse to an identifiable person (care, money or credit, benefits, housing, education, employment, essential services, legal standing, liberty)?

- ☐ Yes → **Tier 2.** Go to Step 4. ☐ No → Step 3.

One sentence of reasoning: _____

Step 3: Dependence question. Do the outputs steer operations that people depend on, without deciding about any identifiable person?

- ☐ Yes → **Tier 3.** ☐ No → **Tier 4.** Go to Step 4.

One sentence of reasoning: _____

Step 4: Modifiers. These move the case up only.

- **Autonomy:** will individual outputs take effect without prior human review? ☐ No ☐ Yes → Tier 3 or 4 moves up one tier; at Tier 2 this is Step 1 territory: STOP and escalate.
If human review is claimed, show the arithmetic: (outputs per week _____ ÷ reviewers _____) ÷ working days per week _____ = _____ reviews per reviewer per day. Is that real review? ☐ Yes ☐ No → treat as no prior human review.
- **Detectability:** if an output were wrong, would anyone notice within days? State the mechanism that would notice it: _____
☐ Yes ☐ No → Tier 3 or 4 moves up one tier; at Tier 2, the strict Module 6 path is mandated (shadow mode before go-live; monitoring designed around unnoticed failures).

Result

Final tier	
Governance load acknowledged (attach or reference the tier table, Module 2.3)	☐
Acceptable-use note written (general-purpose tools only, rule 1)	☐ attached
Re-classification triggers acknowledged (rule 4), plus any system-specific trigger, stated: _____	☐
Next scheduled review date (12 months Tier 2 and 3; 24 months Tier 4)	

Signatures

	Name	Role	Date
Classified by			
Confirmed by (System Owner for Tier 2; a peer outside the proposing team otherwise)			

File this worksheet in the system register (Module 7, Artifact 7.11).

References for this module

- Parlementaire ondervragingscommissie Kinderopvangtoeslag (Parliamentary Interrogation Committee on Childcare Benefits), “Ongekend Onrecht” (“Unprecedented Injustice”), Netherlands House of Representatives, December 2020.
- Autoriteit Persoonsgegevens (Dutch Data Protection Authority), “Boete Belastingdienst kinderopvangtoeslag” (fine of 2.75 million euros on the Minister of Finance for the Tax and Customs Administration’s unlawful and discriminatory processing of applicants’ nationality in the risk classification of childcare benefit claims), published 7 December 2021, autoriteitpersoonsgegevens.nl.
- Amnesty International, “Xenophobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal,” October 2021. (Documents the risk classification model as a black box system including a self-learning algorithm.)
- Facts drawn on: approximately 26,000 parents wrongly accused between 2005 and 2019; a self-learning risk classification model scored claims for fraud risk with officials scrutinising the highest-scored claims; nationality used as a risk factor; resignation of the Dutch cabinet on 15 January 2021.

Module 3: Governance roles and human oversight

Every AI system in your institution needs a small set of named people with real authority over it. This module tells you which four roles to appoint, what each one signs and can refuse, what the hours cost, and how to make human review of AI outputs something more than a signature under the machine's decision.

Appoint the four roles in 3.2; existing staff hold them alongside their jobs, and the definitions in Artifact 3b can go into appointment letters. Then set the four gates in 3.3, design the oversight for each Tier 2 system with the four elements in 3.4, install the countermeasures in 3.5, and put the escalation rules in 3.6 on one page where operators can see them.

3.1 What broken oversight looks like

The sepsis prediction tool in Module 4.1 raised alerts on 18 percent of all hospitalised patients. Every alert nominally had a human in the loop: a clinician who would read it and decide. A ward receiving alerts on nearly one patient in five, most of them false, cannot investigate each one; the rational response, within weeks, is to glance and dismiss. The system's own behaviour had made real review impossible.

The Dutch benefits system in Module 2.1 had review in its design too: officials examined the claims the model scored as high risk. The human step existed on paper, and roughly 26,000 wrongly accused parents experienced what it was worth. Amnesty International found that the officials were given no information on why the system had given a claim its high-risk score.

Both are instances of a pattern that has been measured for decades. **Automation bias** is the human tendency to treat an automated cue as a substitute for looking at the evidence oneself. It produces two error types: **omission errors**, failing to act when the system stays silent even though other information says act, and **commission errors**, following the system's output even when available information contradicts it. In controlled studies, making people accountable for the accuracy of their decisions lowered both error types (Skitka, Mosier, and Burdick, 2000; full citation in the references). Oversight is a design problem with known failure modes and known countermeasures.

In this framework, **oversight** means a named person, with authority they can use, information they can act on, time to look, and accountability for what they decide.

3.2 Appoint the four roles

This is the minimum structure that makes the rest of the framework work in an institution with no compliance department. Four roles are the smallest set that keeps four things out of one pair of hands: accountability for the system, day-to-day operation, control of the data, and challenge from outside the project. The set is this framework's design choice. The EU AI Act does not prescribe it: the Act asks deployers to assign oversight to people with competence, training, and authority, and these four roles are one way of doing that. One person can hold a role for several systems. The same person must never hold both the System Owner and Independent Reviewer roles for the same system.

The System Owner is accountable for the system's behaviour in service. One per system, senior enough to stop the system on their own authority without asking permission. They sign the procurement gate, the go-live gate, and the continued-operation reviews; they see the monthly monitoring summary; they own incident decisions above the operational level; and their name is on the system register entry. When something goes wrong, this is the person who answers for it, which is why the role must sit with someone who had the power to prevent it.

The Operational Lead runs the system day to day. They review the monitoring, keep the acceptable-use note current, receive escalations from operators, hold the authority to pause the system for up to 48 hours on their own decision, and prepare the monthly summary for the System Owner. This is the largest recurring time commitment of the four.

The Data Steward controls what data enters and leaves the system. They complete and sign the Module 5 pre-deployment data checklist, approve any change to the data the system receives or the destinations it sends to, and are the institution's contact for data questions from auditors, regulators, or the vendor.

The Independent Reviewer signs the pre-deployment readiness and go-live gates for Tier 2 systems, as someone with no stake in the system going live. They must sit outside the team deploying the system and have had no role in proposing or procuring it. In a small institution, a peer department head serves; two departments may review for each other, with the arrangement recorded in the register. Their job at the gate is written on the gate document: check that every checklist item claimed as done has a document behind it, and refuse the signature where one is missing.

The cost, in hours. These are framework defaults; adjust them to your setting, and record what you chose on the register. Numbers are per month in steady operation, with one-time costs shown separately.

Role	Steady state, one Tier 2 system	Steady state, Tier 3	Tier 4	One-time (procurement through go-live)
System Owner	4 hours per month	1 hour per month	1 hour per quarter for a portfolio of tools	10 to 15 hours across the gates
Operational Lead	8 hours per month	3 hours per month	Covered by normal IT duties	20 to 30 hours, mostly in validation and rollout (Module 6)
Data Steward	2 hours per month	1 hour per month	One-time checks only	8 to 12 hours, mostly the Module 5 checklist
Independent Reviewer	None (gate work only)	None	None	About 8 hours per gate, twice

Total for one Tier 2 system in steady operation: roughly 14 hours per month across the roles. Leadership approves these hours when it approves the system; the one-page orientation in Module 1 asks for that.

Continuity rules. Roles are held by name in the system register. When a holder leaves or moves, a written handover happens within two weeks: the register is updated, and the incoming holder receives the system's worksheet, limitations statement, and last three monitoring summaries. A Tier 2 system whose System Owner or Operational Lead role has been vacant for more than 30 days (a framework default) fails its next continued-operation gate automatically.

Shared systems. Where one platform serves many institutions (a ministry platform used by hundreds of facilities), each deploying institution names its own System Owner and Operational Lead for its own use, and the platform operator holds a System Owner for the platform itself. Each institution's pause authority covers its own use; only the platform owner can pause the platform.

3.3 Set the four gates

A gate is a point where a named person signs a named document, and the system does not move forward without the signature. Four gates cover the lifecycle. **A gate without a signature is a gate not passed.**

Gate	What is signed	Who signs (Tier 2)	Tier 3	Tier 4
G1: Procurement approval	The Module 4 due diligence checklist, passed at tier threshold, with the risk worksheet attached	System Owner	System Owner	Operational Lead
G2: Pre-deployment readiness	The Module 5 data checklist and the Module 6 validation results against the pre-agreed criteria	Data Steward (data) and Independent Reviewer (whole)	Data Steward	Data checks only
G3: Go-live	The go-live section of the Module 6 lifecycle checklist, including monitoring plan, incident plan, and appeal channel in place	System Owner and Independent Reviewer	System Owner	Operational Lead

Gate	What is signed	Who signs (Tier 2)	Tier 3	Tier 4
G4: Continued operation	The periodic review record: monitoring against thresholds, revalidation result, role continuity	System Owner, every 12 months and after any Module 2 re-tier trigger	System Owner, every 12 months	Every 24 months

Each signed gate document is filed in the system register (Module 7). The gates are where the four roles' authority becomes visible: G2 and G3 are the places where the Independent Reviewer's refusal, or the Data Steward's unfinished checklist, stops a launch that momentum would otherwise carry. If leadership decides to proceed over an Independent Reviewer's refusal, it may do so only in writing, recorded in the register beside the refusal, as a risk acceptance signed at leadership level.

3.4 Make oversight real

For every Tier 2 system, design the human review around these four elements, in writing, before go-live. The design is one page; Artifact 3b's Operational Lead definition says who drafts it and the G3 gate is where it is checked. A review arrangement missing any one of the four is nominal, and section 3.5 tells you how to catch it.

Authority. The reviewer can overrule the system's output in the individual case, and can trigger a pause of the system, without asking anyone's permission first. Both powers are written in their role or job description, and the override is executable in the tools they use: if rejecting the system's recommendation takes four approvals and a form the software does not provide, the authority is fictional whatever the paper says. Test it before go-live: have a reviewer override a live-like case end to end, and time it.

Capacity is training plus time, and both are measurable.

- **Training:** every reviewer has read the system's limitations statement (the document Module 4.2.3 extracted from the vendor) and has seen worked examples of the system being wrong, drawn from validation. Refresh when the model updates (Module 6).
- **Time:** during shadow mode (Module 6), measure how many minutes a review of one case takes. That number becomes the review budget. Then do the division that the Module 2 worksheet already forced: expected outputs per week, divided by reviewers, against the budget. If the arithmetic does not fit inside real working hours, the design changes now: more reviewers, a triage rule that routes only some outputs to full review (for example: full review for every adverse recommendation, sampling at a stated percentage for the rest, with the split written in the oversight design), or a narrower deployment. What must never close the gap is silently shrinking the minutes.

Information. The reviewer sees, at minimum: the system's output; the inputs it was based on; whatever the system can show about its reasoning or confidence (the answer to vendor question 11 in Module 4.4); and the underlying record itself, one click away. A reviewer shown only a score and an approve button has been staged for commission errors, because the design has removed every piece of information that could contradict the system.

Consequence. Decisions leave a trace, and the trace is looked at. For adverse Tier 2 decisions, the reviewer records a one-line reason when approving the system's recommendation, in their own words. This is the accountability that the research says lowers both error types, and it costs fifteen seconds per case. Overrides are recorded with reasons too, and the monthly monitoring review reads both streams. The protective half of consequence is stated in writing: no one is penalised for overriding the system or triggering a pause in good faith, and if a manager reverses a reviewer's override, that reversal is itself recorded and lands in the System Owner's monthly summary.

By tier: Tier 2 gets all four elements, designed and gate-checked as above. Tier 3 gets authority and information (a named person who can pause, seeing forecasts beside actuals). Tier 4 gets the acceptable-use note and a named Operational Lead, which is oversight enough for tools whose errors cost rework.

3.5 The rubber-stamp trap, and the countermeasures

A reviewer approving everything looks, from every dashboard, like a system that is always right. These countermeasures make the difference visible, and each comes with a number.

1. Track override rates, and treat quiet as a signal. The monthly monitoring summary (Module 6) reports, per system and per reviewer: cases reviewed, approvals, overrides, and the reason categories. Investigate when the override rate stays below 1 percent for a quarter at meaningful volume, framework default: 500 or more cases in the quarter (a thousand-case quarter with three overrides means the humans have stopped looking or the system has become perfect). Investigate equally when it exceeds 30 percent (the system and its reviewers disagree so often that either the system is misfiring or trust has broken, and both are findings). Both percentages are framework defaults: set your two thresholds at go-live and write them on the monitoring plan.

2. Blind second review of a sample. Each month, a second reviewer re-decides a sample of cases without seeing the system's output or the first reviewer's decision: 2 percent of cases or 10 cases, whichever is larger. Compare. Disagreement between the blind reviewer and the recorded decisions is the direct measurement of what the system-plus-reviewer pipeline is doing to outcomes, and four hours a month of a second reviewer's time is what the measurement costs.

3. Seeded test cases, where the workflow permits. Quarterly, insert a small number of known cases whose correct handling is already established (synthetic records in queue-based workflows such as claims, applications, or document review; never in workflows where a test case could touch a real person's care or file). A reviewer who waves through a seeded clear failure gets the retraining in 3.4; a second miss moves them off review duty for that system. Tell reviewers seeding exists and never when. Where seeding is impracticable, say so in the oversight design and lean harder on countermeasure 2. **Check locally:** reviewer performance measures in this section, seeding included, must comply with local employment law and any staff representation or collective agreements; the existence of seeding is disclosed policy rather than covert surveillance, and only its timing is undisclosed.

4. Written reasons for approvals, per 3.4's consequence element, on adverse Tier 2 decisions. Reasons of "as per system" count as no reason, and the Operational Lead's monthly read of a reason sample is on the RACI.

5. Rotate. No single reviewer owns a Tier 2 review queue alone for more than 12 months without rotation or a second chair.

The five countermeasures cost, together, on the order of a working day per month for a Tier 2 system.

3.6 Escalation, and the pause rule

Escalation is a ladder with times on it, published to everyone who touches the system, one page, posted where operators see it. The framework default:

Step	Who	Time
Any operator, reviewer, or appeal channel raises a concern	To the Operational Lead, same working day	Hour zero
Operational Lead assesses; may pause the system on own authority (up to 48 hours)	Decision to pause or continue, recorded with one line of reasoning	Within 4 working hours
System Owner informed of every pause and every Severity 1 or 2 concern (Severity 1: wrong or harmful outputs are reaching people now; Severity 2: the system is unavailable or known to be producing unreliable outputs but they are contained; definitions in the incident list, Module 6.6)	Confirms, extends, or lifts the pause; declares an incident if the Module 6 definition is met	Within 24 hours
Incident path (Module 6) takes over if declared	Containment, assessment, notification per the incident template	Per the Module 6 clock

The pause rule: anyone who works with the system can request a pause review, the request cannot be refused silently (the Operational Lead's 4-hour decision and one-line reasoning are mandatory), and no one is penalised

for a good-faith request, in writing, in the oversight design. **The bypass rule:** if a pause or an override is bypassed or reversed by anyone, the bypass is recorded and appears in the System Owner's next summary; a bypass whose output reached an affected person is an incident under the Module 6 definition. **The Severity 1 rule:** where a system's output is about to reach a person and a reviewer believes it is wrong or harmful, stopping that single output is within every reviewer's existing authority under 3.4, and needs no ladder at all.

Where a vendor-hosted system offers no pause control, the pause is executed on your side: suspend the institution's use of the outputs, route the work back to the manual process, and instruct the vendor under the support clause (Module 4.3.12). And for systems supporting emergency operations, the pause rule operates through the fallback: the Module 6 rule that no high-stakes system stops into a vacuum applies to pauses too, so the oversight design names the fallback that takes the load the moment a pause lands.

3.7 Where this module sits in existing standards

This module	Where it maps
Named roles, decision rights, gates	The NIST AI RMF Govern function: accountability structures, defined roles and responsibilities, and organisational culture around AI risk. ISO/IEC 42001's leadership and roles clauses: management responsibility assigned and documented within the management system.
3.4's four elements of real oversight	The EU AI Act's human oversight requirements for high-risk systems: oversight by persons able to understand the system's capacities and limitations, to interpret outputs, to decide not to use them, and to intervene or interrupt. The Act also requires deployers to assign oversight to people with the training and authority to carry it out. Precise article references in the annex.
3.5's countermeasures	The AI RMF Measure and Manage functions applied to the human layer: the oversight arrangement is itself monitored, measured, and corrected.
The pause rule and recorded reversals	The accountability principle in the OECD AI Principles and the UNESCO Recommendation, and UNESCO's insistence on human determination: a human can always decide, and the record shows who did.

Artifact 3a: Governance RACI

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0. R = Responsible (does the work). A = Accountable (answers for it). C = Consulted. I = Informed. Adjust to your structure and file the adjusted version in the system register. Tier note: this chart shows the Tier 2 allocation. At Tier 3, the Independent Reviewer column falls away (Module 3.2). At Tier 4, the Operational Lead holds the A for the G1 and G3 gate rows, per the gate table in Module 3.3.

Activity	System Owner	Operational Lead	Data Steward	Independent Reviewer
Risk classification worksheet (Module 2)	A (confirms Tier 2)	R	C	C (confirms below Tier 2, if peer)
Tender demands and vendor interview (Module 4)	A	R	C (data demands)	I
Due diligence checklist and G1 signature	A	R	C	I
Contract signature readiness (clauses, counsel review)	A	C	C	I
Pre-deployment data checklist (Module 5)	I	C	A and R	I
Validation on own data (Module 6)	A	R	C	I
G2 readiness signature	I	C	R (data part)	A (whole)
Oversight design (3.4 one-pager)	A	R	C	C (checks at G3)
G3 go-live signature	A	R (evidence pack)	C	A (joint)
Monthly monitoring review and summary (Module 6)	I (receives)	A and R	C (data metrics)	I
Override, reason, and blind-sample review (3.5)	I	A and R	I	I
Pause decisions (3.6)	A (beyond 48 hours)	R (first 48 hours)	I	I
Incident declaration and response (Module 6)	A	R	C (data incidents: R)	I
Vendor update acceptance and re-test (Module 6)	A	R	C	I
Grievance and appeal responses (Module 6)	A	R	C	I
G4 continued-operation review	A	R	C	I
Decommissioning and data return (Modules 4 and 6)	A	R	R (deletion certificate)	I

Artifact 3b: Role definition pack

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0. Four half-page definitions, written to be pasted into an appointment letter or job description. Replace bracketed items. Time budgets are the framework defaults from Module 3.2; adjust and record. Tier note: at Tier 4, the G1 and G3 gate signatures sit with the Operational Lead (Module 3.3); adjust the two definitions accordingly when appointing for a Tier 4 portfolio.

System Owner, for [system name]

Purpose. You are accountable for this system's behaviour in service. You are the person who answers for it, to leadership, to auditors, and to affected people, and you hold the authority that accountability requires. **Authority.** You may halt the system at any time on your own decision. You approve or refuse the procurement gate (G1), the go-live gate (G3, jointly with the Independent Reviewer), and the periodic continued-operation gate (G4). You own incident decisions beyond the Operational Lead's 48-hour pause authority. **You sign.** The due diligence checklist (G1), the go-live checklist (G3), continued-operation reviews (G4), and the register entry carrying your name. **You may refuse.** Any gate where a claimed item lacks a document. Refusals are recorded, and recorded refusal is a protected act. **You receive.** The monthly monitoring summary, every pause and bypass record, and every incident declaration. **Time.** About 4 hours per month in steady operation for a Tier 2 system; 10 to 15 hours one-time across procurement and go-live. **Seniority test.** If you would need permission to stop this system, the role sits too low. Hand it up.

Operational Lead, for [system name]

Purpose. You run this system day to day and you are the first pair of eyes on everything it does in service. **Authority.** You may pause the system for up to 48 hours on your own decision. You decide pause requests within 4 working hours, with one recorded line of reasoning. You accept or defer vendor updates into the test window (Module 6) and keep the acceptable-use note current. **You do.** The monitoring review and the monthly summary; the override-rate, written-reason, and blind-sample countermeasures (Module 3.5); reviewer training records; the oversight design one-pager (Module 3.4) for the Independent Reviewer's G3 check; escalation intake. **You may refuse.** To lift a pause you imposed until the System Owner has reviewed it. **Time.** About 8 hours per month in steady operation for a Tier 2 system; 20 to 30 hours one-time through validation and rollout.

Data Steward, for [system name]

Purpose. You control what data this system receives, produces, and sends onward, and nothing about that changes without you. **Authority.** You approve or refuse the pre-deployment data checklist (Module 5), and any subsequent change to data sources, fields, destinations, or storage locations. Your refusal blocks G2. **You sign.** The data section of G2; deletion certificates at decommissioning. **You are the contact for.** Data questions from auditors, regulators, affected people exercising data rights (check locally which rights apply), and the vendor. **Time.** About 2 hours per month in steady operation; 8 to 12 hours one-time, mostly the Module 5 checklist.

Independent Reviewer, for [system name]

Purpose. You are the outside check with the right to refuse the signature, at the two gates where enthusiasm and momentum are least reliable: readiness (G2) and go-live (G3). **Independence.** You sit outside the team deploying this system and had no part in proposing or procuring it. If your department and theirs review for each other, the arrangement is recorded in the register. **Your method, per gate.** Take the gate checklist. For every item marked done, ask for the document. Read the validation results against the criteria that were agreed before testing, and check the oversight design one-pager against the four elements in Module 3.4. Where a document is missing or a criterion unmet, refuse the signature and record why. You are not asked to judge the technology; you are asked to confirm the evidence exists. **You may refuse.** The signature. That is the role. **Time.** About 8 hours per gate, twice per system.

References for this module

- Skitka, L. J., Mosier, K. L., and Burdick, M. "Accountability and automation bias." *International Journal of Human-Computer Studies*, volume 52, issue 4 (2000), pages 701 to 717. doi:10.1006/ijhc.1999.0349.
- Wong, A., et al. "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients." *JAMA Internal Medicine*, volume 181, issue 8 (2021), pages 1065 to 1070. (Full citation in Module 4; cited here for the alert volume of 18 percent of hospitalised patients.)
- Case facts on the Dutch childcare benefits affair as cited in Module 2.

Module 4: Procurement and vendor accountability

The moment of signature is the only moment when your institution has more power than its AI vendor. This module tells you what to demand in the tender, what to verify in the vendor interview, and what to lock into the contract before that moment passes.

Read sections 4.1 to 4.7 once, in order, before your next AI purchase. Then work from the two artifacts at the end: the vendor due diligence checklist (Artifact 4a) and the contract clause library (Artifact 4b). If a purchase is already in progress, start with the red flags in section 4.5 and the checklist.

This module is an operating manual and it is not legal advice. Contract law is national law. Before your institution signs anything based on this module, the final contract text must be reviewed by a lawyer qualified in your jurisdiction. Section 4.3 tells you where that review matters most.

4.1 Why procurement is the control point

In 2021, researchers at a large university hospital in the United States published an independent test of a sepsis prediction tool. Sepsis is a life-threatening reaction to infection, and catching it early saves lives. The tool came built into the hospital's electronic health record software and was already in use at hundreds of hospitals. The developer's reported accuracy, as recorded in the study and in the JAMA Internal Medicine editorial that accompanied it, was an area under the curve of 0.76 to 0.83 (area under the curve, or AUC, is a standard accuracy score where 1.0 is perfect and 0.5 is a coin flip).

The researchers ran the tool against 38,455 admissions at their own hospital. On their patients, the measured AUC was 0.63. The tool failed to flag 67 percent of the patients who developed sepsis. At the same time it raised alerts on 18 percent of all admitted patients, so many that clinicians could not treat the alerts as meaningful. The study is Wong and colleagues, "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients," JAMA Internal Medicine, 2021 (full citation in the references at the end of this module).

The tool was already purchased, installed, and running in hundreds of institutions. The gap between the claimed performance and the measured performance was discoverable before any of those purchases. It required a question in the tender ("on what population were these accuracy figures measured?") and a test on the buyer's own records before go-live. Neither requires a data science team.

This module exists because of an asymmetry. The vendor has built, sold, and supported this system many times. You are buying it once. The vendor knows its failure modes, its real performance, and the weaknesses in its own standard contract. On the day you sign, you know what the sales material told you. There is one force that closes this gap: before signature, the vendor wants something from you. Every protection in this module costs one paragraph in a tender or one clause in a contract while the vendor is still selling. After signature, the same protection costs a negotiation, and often a fee, and sometimes it cannot be had at any price.

A few terms, used throughout this module:

- **Procurement** is the whole process of buying: defining what you need, inviting offers, evaluating them, negotiating, and signing.
- **A tender** (in some countries: a request for proposals, or RFP) is the written document you send to vendors describing what you require. Everything you demand in the tender, vendors must respond to in writing, and their written responses can be made part of the contract.
- **A vendor** is the company selling you the system. Where the seller is a reseller or integrator rather than the developer, the demands in this module apply to whoever can answer them, and a reseller who cannot pass your questions through to the developer is itself a finding.
- **Due diligence** is the structured checking you do before committing.

The order of work in a purchase:

1. **Write the demands into the tender.** Section 4.2 lists them. Vendors respond in writing.

2. **Score the responses.** Artifact 4a is the scoring checklist. It has pass thresholds by risk tier (risk tiers are set in Module 2).
3. **Interview the shortlisted vendors.** Section 4.4 is the question script. You are testing whether the written answers survive a live conversation.
4. **Negotiate the contract.** Section 4.3 explains the terms that matter and the floor below which you do not go. Artifact 4b is the clause library.
5. **Sign only when the checklist passes** for the system's tier, and route the signed checklist through the procurement approval gate defined in Module 3.

This module governs the buying. Module 6 governs the running. Acceptance testing, the practice of testing the system on your own data before you rely on it, appears in both: here as a right you write into the contract, there as a stage you execute before go-live.

Check locally: some public procurement regimes leave no room to negotiate after tenders close, and some prescribe standard terms. That makes this module more useful there, and differently used: in those regimes the tender is the negotiation, because every demand from 4.2 and every clause requirement from 4.3 written into the tender becomes binding through the vendor's response, without a bargaining table. Where your procurement law fixes a term you cannot change, record the gap on the due diligence checklist as a risk the institution accepts knowingly, and check whether your regime's standard clauses already cover it.

4.2 What to demand before signing

If a vendor's claim matters to your decision, get it in writing, and make the writing part of the contract.

Everything below goes into the tender as a required deliverable. Each demand here reappears as a line in the due diligence checklist (Artifact 4a) and most have a matching clause in the library (Artifact 4b).

4.2.1 System and purpose documentation

What to demand: a system description document, in plain language, covering: what the system does; its intended uses and intended users; its explicitly excluded uses; what data goes in and what comes out; whether it is a predictive system (it scores, ranks, classifies, or forecasts) or a generative system (it produces text, images, code, or other content); and a general description of the data it was trained on, including the time period, the countries or regions, the languages, and the kinds of people or records it covers.

The excluded-uses list is where vendors tell you where the system breaks. The training data description is what lets you answer the population question in 4.2.2. You do not need the training data itself, and most vendors will not provide it. You need to know what world the system learned from, so you can judge how far your world is from it.

Tender language you can lift: "Provide a system description document covering intended use, excluded uses, inputs, outputs, system type, and a general description of training data including time period, geography, languages, and population coverage. This document will be incorporated into the contract."

4.2.2 Performance evidence, and the population question

What to demand: every performance claim the vendor makes, in writing, with three things attached to each claim: the metric used, the dataset or environment it was measured on, and a description of the population in that dataset. Then the direct question, in the tender, in these words or your own: **"Was this system's performance measured on people like the people we serve? Provide the evidence."**

For any system whose outputs affect people (Tier 2 in Module 2), also demand performance broken out by subgroup, chosen by a simple rule: the groups your institution already reports on elsewhere, plus any group your national law protects. If your institution reports nothing by group today, start with age, sex, and language. Ask what the vendor measured, and ask what they did not measure.

Why the population question exists. In 2019, researchers published an analysis in the journal *Science* of a commercial algorithm used across the United States health system to decide which patients should be offered

extra care management. The algorithm affected millions of patients. It predicted each patient's future healthcare costs and treated predicted cost as a measure of medical need. Because less money had historically been spent on the care of Black patients than on equally sick White patients, the algorithm concluded that Black patients were healthier than they were. At the same risk score, Black patients were considerably sicker. The researchers calculated that removing this bias would raise the share of Black patients identified for extra help from 17.7 percent to 46.5 percent. The study is Obermeyer, Powers, Vogeli, and Mullainathan, "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, 2019 (full citation in the references). No buyer could have caught this by reading accuracy figures. The overall accuracy looked fine. It could have been caught, and was eventually caught, by asking what the system predicts, and how its outputs distribute across groups of real people. That is why 4.2.1 demands a description of what the system does in plain terms, and why this section demands subgroup evidence rather than a single headline number.

Also demand: the vendor's method, in enough detail that the result could be checked. You may never check it yourself. Demanding it changes what the vendor sends you.

4.2.3 Known limitations and failure modes

What to demand: a written limitations statement: the conditions under which performance degrades, the kinds of inputs the system handles badly, known failure modes, and any populations, languages, or settings where it has not been tested.

Every real system has limitations and every serious developer has documented them internally. A vendor who provides none is telling you one of two things: they do not know their own system, or they will not tell you. Either answer disqualifies the vendor for a Tier 2 purchase. Module 3 requires the humans who oversee the system to be trained on its known failure modes, and this statement is the training material.

4.2.4 Audit logs

What to demand: written confirmation of what the system records about its own operation. An **audit log** is the system's own record of what happened: what input was received, what output was produced, when, under which model version, and at whose request. Demand: the list of what is logged; who can access the logs; how long they are kept; and that your institution can export them in a documented format.

Without logs you cannot answer a complaint from an affected person (Module 6), investigate an incident, or check a vendor's account of events against the record. A system that keeps no logs, or whose logs only the vendor can read, cannot be governed. For Tier 2 systems, treat missing or vendor-locked logs as disqualifying.

4.2.5 Security and data handling attestations

What to demand, at tender stage: where your data will be stored, processed, and backed up, by country; who at the vendor and its subcontractors can access it; whether your data or your users' inputs will be used to train or improve any model, now or later; and copies of any security certifications the vendor claims, together with the scope statement of each certificate (a certificate can cover one product, one office, or one process; the scope statement tells you which).

Module 5 is where your institution settles its own data questions before anything leaves the building. This section is the vendor-facing half: the answers you must hold in writing before you can even run the Module 5 checklist. Demand them now, while the tender gives you the power to demand.

4.2.6 Incident history and notification commitment

What to demand: a summary of security and performance incidents affecting this product over a defined period (three years is a workable ask), and the vendor's standard commitment for notifying customers of incidents, expressed as a number of hours or days. Vagueness here ("promptly," "without undue delay") is a finding: it means the vendor has no internal clock, or has one and will not show you.

4.2.7 Update and change policy

What to demand: the vendor's written policy on changing the system in service. Specifically: whether the model can change without notice to you (a **silent update**); how much advance notice you get; whether changes come with a changelog (a written list of what changed); and whether you can test a change before it reaches your live system, or decline it for a period.

The system you validate before go-live is only the system you run until the first update. Module 6 defines the operational routine for handling updates. That routine is impossible unless the rights are in the contract, and the rights are only in the contract if the demand was in the tender.

4.2.8 Additional demands for generative tools

Generative systems, including chat assistants, drafting aids, and code assistants, need everything above plus four demands of their own:

1. **Error rates, measured.** Generative systems produce confident false statements (commonly called **hallucinations**). Demand the vendor's measured rate of factual error, and the method behind the measurement. A vendor who says the rate is zero, or that it cannot be measured, has answered your question.
2. **Input attack testing.** A **prompt injection** is an attack where instructions hidden in the text a system reads (an email, a document, a web page) hijack its behaviour. Demand written evidence that the product has been tested against this class of attack, and what the results were.
3. **Retention and training on your inputs.** Demand, in writing: are user prompts and uploaded content retained, for how long, and are they used for training? For institutional deployment the answer you require is a configuration where your users' inputs are excluded from training. If that configuration does not exist, route the finding into Module 5 before any real data touches the tool.
4. **Per-user logging.** Demand that the tool can log which user submitted what and received what, exportable by your institution. Without this, you cannot investigate misuse or answer a complaint.

Check locally: where the EU AI Act applies, its Article 50 transparency duties apply from 2 August 2026. Ask the vendor, in writing, how the product tells people they are dealing with an AI system and how it marks the content it generates. Your institution must disclose deepfakes, and AI-generated text published to inform the public on matters of public interest, unless the text has had human review and someone holds editorial responsibility for it.

A note on scale: these four demands apply even when the tool is cheap or free, because free tools take payment in data. Section 4.6 gives the shortened path for low-tier purchases.

4.3 The contract terms that matter

This section walks through the terms in plain language. For every term except two, Artifact 4b gives model wording (the exceptions: liability, 4.3.9, which is deliberately left to your counsel, and the support severity definitions, 4.3.12, which are given as pasteable text in the section itself), and the marker **"Do not concede"** states the floor for Tier 2 systems: the point below which you walk away rather than sign.

The clause library is drafting guidance. Contract law is national law, and it varies. A clause that protects you in one jurisdiction may be unenforceable in another. Before signature, the final contract must be reviewed by a lawyer qualified in your jurisdiction. Sections 4.3.9 (liability) and 4.3.7 (breach notification) are where local law matters most; say so to your lawyer and they will know where to look.

4.3.1 Documentation becomes warranty

Every document the vendor produced under section 4.2, and every written performance claim, gets attached to the contract and stated to be accurate. If a claim later proves materially false, that is a breach of contract with remedies, on the same footing as a failure to deliver.

It converts the sales process into commitments. **Do not concede:** a contract that states the vendor's marketing and pre-contract statements have no legal effect (a common clause called "entire agreement" often does this) while refusing to attach the documentation. If the pre-contract claims are excluded, the claims you relied on to buy do not exist.

4.3.2 Acceptance testing and a performance warranty

The contract states: before the system goes live, and before the main payment milestone, your institution will test it on its own data against criteria agreed in the contract, and the system must meet them. (How to run that test without a data science team is Module 6, section 6.2. Here you are securing the right, the vendor's cooperation, and the consequence of failure.) The contract also states an ongoing standard: the performance the system must maintain in service, and the agreed way of measuring it.

The sepsis case in 4.1 becomes contractually impossible to ignore. A system that scores 0.63 where 0.76 was promised fails acceptance, and failing acceptance has a defined consequence: fix, renegotiate, or exit with your money. **Do not concede:** a contract where the only performance standard is the vendor's own benchmark on the vendor's own data. That standard was already met before you appeared.

4.3.3 Audit rights

Your institution has the right, at least annually and additionally after any incident, to audit the system's logs, the vendor's compliance with the contract's data and security commitments, and the documentation, either directly or through an independent third party bound by confidentiality.

The vendor's commitments stay true after the sales team has gone. **Do not concede:** for a Tier 2 system, no audit rights means no purchase.

4.3.4 Update and change control

No change to the model or to system behaviour reaches your live environment without: advance written notice (put a number of days in the contract; 30 is a common and workable ask for planned changes), a changelog, a window in which your institution can test the change, and the right to remain on your current version for a defined period if the test fails. Emergency security fixes are the exception, with notice to follow promptly after.

It makes Module 6's revalidation routine possible. Without this clause, the vendor can silently replace the system you validated with one you have never seen. **Do not concede:** silent updates to a Tier 2 system.

4.3.5 Data rights

Your institution's data remains your institution's. The vendor may use it only to provide the contracted service. The vendor may not use it to train models, improve products, or build anything else without a separate written agreement signed by your institution. On exit, the vendor deletes your data from its systems and confirms the deletion in writing.

Do not concede: any clause granting the vendor broad rights to "aggregate," "anonymise and use," or "derive insights from" your data by default. If the vendor wants your data for its own purposes, that is a separate negotiation with its own price, made with your eyes open and Module 5 completed.

4.3.6 Data location and subcontractor transparency

The contract names the countries where your data is stored, processed, and backed up, and names the subcontractors who touch the service. Changes to either require advance notice and, for Tier 2 systems, your written consent. Every obligation the vendor owes you flows down to its subcontractors.

Without it, the vendor's promises stop at the vendor, and the subcontractor holding your data owes you nothing. **Check locally:** many jurisdictions restrict cross-border transfer of personal or government data. Module 5, section 5.3 gives you the questions to settle with local advice; this clause is where the answers get enforced.

4.3.7 Security requirements and breach notification

The security controls the vendor committed to in the tender go into the contract as a schedule (an attached list), and the vendor commits to notify your institution of any security breach affecting your data within a stated number of hours. Write a number. Commitments in real contracts commonly range from 24 to 72 hours; pick the strictest the vendor will sign. **Check locally:** many jurisdictions impose legal deadlines for notifying regulators and affected persons after a breach. Your vendor's clock must run faster than your legal clock, because you cannot notify anyone about what you have not been told.

4.3.8 Incident notification beyond breaches

Security breaches are one kind of incident. The clause you also need: if the vendor discovers a material defect, error, or safety problem in this product, including one discovered at another customer, it notifies you within a stated number of days. The sepsis tool from section 4.1 ran in hundreds of hospitals; a defect found in one is a defect running in all of them.

4.3.9 Liability

Two mechanisms in nearly every vendor contract decide who pays when things go wrong. First, a **liability cap:** a ceiling on what the vendor can ever owe you, often expressed as a multiple of the fees you paid. Second, **exclusions:** categories of loss the vendor never pays, commonly labelled "indirect" or "consequential" loss, labels which can be read to swallow most of the harm that would matter to you.

What you can do without a lawyer: find the cap, express it as a number, and put it beside a realistic bad outcome for this system at your institution. A cap of one year's fees on a system that could misdirect care, benefits, or credit decisions for thousands of people is a number you can evaluate yourself. Then list the exclusions and check whether breach of the data clauses (4.3.5, 4.3.7) sits inside the cap or outside it; vendor liability for data breaches and for violating your data rights should sit outside the general cap or under a separate, higher one.

Check locally: what these words mean, and whether they hold, is decided by national law and courts. This is the clause your local lawyer earns their fee on. Bring them the cap number, the exclusion list, and your realistic bad outcome, and ask one question: if this happens, what do we actually recover?

4.3.10 Exit and portability

The contract guarantees that on termination for any reason: the vendor exports all your data, your configurations, and your audit logs in open, documented formats within a stated number of days, at a cost fixed in the contract; the vendor provides transition assistance for a stated period; and these export obligations survive even if the termination is disputed.

It makes leaving possible, which is what makes every other clause enforceable. **Do not concede:** never sign a contract you cannot exit.

4.3.11 Termination triggers

The contract names the events that give your institution the right to terminate: the system fails acceptance testing (4.3.2); the system fails a periodic revalidation under the agreed measurement method; the vendor makes an unnotified change in breach of 4.3.4; the vendor refuses an audit under 4.3.3; or the vendor materially breaches the data clauses. Naming the triggers matters because "material breach," the default standard in many contracts, is an argument, while a named trigger is a fact.

4.3.12 Support and escalation

Response and resolution time targets, by severity level, with a named support channel and a named escalation contact at the vendor. Severity definitions belong in the contract too, or the vendor decides after the fact that your outage was minor. A usable three-level definition you can paste into the schedule: Severity 1, the System is

producing wrong or harmful outputs in live use, response within [x] hours; Severity 2, the System is unavailable, response within [x] hours; Severity 3, degraded performance or a cosmetic fault, response within [x] working days.

4.3.13 Intellectual property indemnity, for generative systems

A generative system produces content, and content can infringe: output that reproduces copyrighted material the model absorbed in training exposes the institution that published the output, not only the vendor that trained the model. The term: the vendor defends and indemnifies the institution against third-party claims that the system, or output produced by the system when used as documented, infringes intellectual property rights, and at its own cost secures continued use, modifies the system, or refunds and assists exit. Clause 13 gives the wording. **Do not concede**, for any generative system whose output leaves the institution: a contract that leaves your institution answering for what the vendor's training data contained.

4.3.14 Price protection and product continuity

AI pricing and product lines change fast. Two sentences for the contract: renewal price increases are capped at [x] percent or an agreed index; and the vendor gives [x] months' notice of withdrawal or end of support of the product or the model version relied on, during which exit assistance under the exit clause is available at contracted rates. Clause 14 gives the wording; Clause 11's termination triggers and Clause 12's continuity arrangement carry the enforcement.

4.4 Questions that expose a weak vendor

Use this script in a live meeting with the shortlisted vendors, after you have their written tender responses. You are testing two things: whether the written answers survive a conversation, and how the vendor behaves when asked. A strong vendor welcomes these questions. They have the answers because they did the work.

Practical mechanics: two people from your institution attend, one asks, one writes down the answers verbatim. The written record goes into the procurement file, and under section 4.3.1 the answers become contractual. Tell the vendor this at the start of the meeting.

For each question below: what a strong answer sounds like, and what a weak one tells you.

1. **"What populations, in what countries and languages, was this system's performance measured on?"** Strong: a specific answer with a document behind it. Weak: "it works globally," "our customer base is very diverse," or any answer describing where it is sold rather than where it was measured.
2. **"Which of your published performance figures would you expect us to reproduce on our own data, and why?"** Strong: a direct engagement with the difference between their test population and yours, and support for an acceptance test. Weak: hesitation at the idea that you would measure at all.
3. **"Show us your documented limitations for this product."** Strong: a real document, produced without friction. Weak: "there are no significant limitations." Every real system has them. This answer means they do not know or will not say.
4. **"What happens when the system receives an input unlike anything in its training data?"** Strong: a concrete description of behaviour: a confidence flag, a refusal, a routing to a human. Weak: "that doesn't really happen."
5. **"What does the system log, who can read the logs, and how do we export them?"** Strong: a list, a retention period, and an export format. Weak: "logging is handled on our side," which means the record of your own system's decisions belongs to someone else.
6. **"Who at your company, and at which subcontractors, can access our data, and from which countries?"** Strong: a named, current answer, because access control means someone maintains this list. Weak: needing to find out, for a vendor already holding other institutions' data.
7. **"Is our data, or our users' inputs, used to train or improve any model, and can that be switched off in writing?"** Strong: a clear default and a contractual toggle. Weak: distinguishing between "training" and "improving services" in a way that preserves their access to your data.

8. “Walk us through your last significant incident on this product: what happened, when did customers learn of it, what changed afterwards?” Strong: a real account, told plainly. Every mature product has incidents; the account tells you how they will treat you when it is your incident. Weak: “we have never had one.”

9. “When you update the model, how much notice do we get, what changelog, and can we test before it reaches our live system?” Strong: an existing process with named notice periods. Weak: “updates are seamless and automatic.”

10. “Give us two reference customers of our type and size, and permission to ask them what went wrong in the first year.” Strong: names, and comfort with the second half of the question. Weak: references only from customers unlike you, or references you may not ask about problems.

11. “For an individual output, what can the system show a reviewer about why it produced that output?” Strong: a concrete demonstration: contributing factors, a confidence score, source passages, whatever the honest answer is for this system type, including “little, and here is how our customers oversee it anyway.” Weak: the word “proprietary” doing the work an answer should do.

12. “What performance differences have you measured across subgroups, and what did you do about them?” Strong: they measured, found differences (real measurement finds differences), and can describe the response. Weak: “the system does not consider protected attributes, so it cannot be biased.” The care management case in section 4.2.2 involved a system that never used race as an input.

13. “If we terminate, what do we get back, in what format, in how many days, and at what cost?” Strong: numbers, formats, a fixed price. Weak: “we’ve never had a customer leave.”

14. For generative tools: “What is your measured factual-error rate, and how did you measure it?” Strong: a number, a method, and honesty about its limits. Weak: “hallucinations are largely solved” or a claim that measurement is not possible.

15. For generative tools: “Show us the results of your prompt injection testing.” Strong: evidence of testing this attack class, and a patch history. Weak: unfamiliarity with the term, for a company selling a system that reads untrusted text.

16. “Which of the answers you have given us today will you attach to the contract?” Strong: “all of them.” This is the question that separates a sales conversation from a commitment. Ask it last, and in those words.

When they cannot answer. Distinguish two situations. “We will confirm that in writing within a week” is acceptable once or twice; complex products have specialists, and the written follow-up becomes part of the record. What is disqualifying is deflection: “proprietary” as the answer to a safety question, an answer to a different question than the one asked, or surprise that you asked at all. A vendor who cannot answer basic questions about its own system in a sales meeting, with the sale on the line, will not answer them in year two, when nothing is on the line.

4.5 Red flags that stop a purchase

Each item below has appeared already; this list exists so that a busy reader can check a live purchase against it in five minutes. These are patterns that predict a bad outcome strongly enough to justify stopping.

1. **Refusal of audit rights.** There is no benign reading of this for a Tier 2 system.
2. **Performance claims with no stated evaluation population.**
3. **“Proprietary” as the answer to safety, performance, or data questions.** Protecting source code is legitimate. Refusing to say how a system was tested is a different act using the same word.
4. **No documented limitations.** See question 3 above.
5. **A silent-update policy, defended as normal.** You would be buying a system that can become a different system without your knowledge.
6. **No data deletion commitment on exit.** Your data outlives the relationship, in someone else’s hands.
7. **No export capability, or export in closed formats only.** This is a lock-in mechanism, priced into the renewal negotiations you will face later.

8. **Pressure to sign before validation.** “You can test after go-live” reverses the point of acceptance testing.
9. **Contract terms that live outside the contract.** An order form that incorporates online policies the vendor can change unilaterally means you are signing terms that do not exist yet. Demand the full terms as a fixed document.

The stopping rule: one red flag pauses the purchase until it is resolved in writing. Two or more red flags on a Tier 2 purchase end the process with that vendor. Record the decision and the reason in the procurement file. A written record of a stopped purchase protects the person who stopped it.

4.6 Scaling this module to the risk tier

Applied in full, this module suits Tier 2 purchases: systems whose outputs affect decisions about people (Module 2 defines the tiers). Applying it in full to every purchase would teach your institution to bypass it.

Tier 2 (high stakes): the full module. All demands in 4.2, all clauses in 4.3, the interview in 4.4, checklist pass at the Tier 2 threshold (Artifact 4a).

Tier 3 (operational): the documentation core plus six clauses. Demand items 4.2.1 through 4.2.5 and, for generative tools, 4.2.8. Contract clauses: documentation as warranty (4.3.1), audit rights (4.3.3), update notice (4.3.4), data rights (4.3.5), breach notification (4.3.7), exit and portability (4.3.10). Interview optional; checklist pass at the Tier 3 threshold.

Tier 4 (low stakes): the short path. Even for a cheap or free drafting aid, three questions and three clauses:

- Questions: Is our data or our users' input used for training, and can it be switched off? What is retained and for how long? Can we export and delete our data on exit?
- Clauses (or, for off-the-shelf tools, settings and written terms you verify before rollout): data rights (4.3.5), update notice (4.3.4), exit (4.3.10).

A note on free and consumer-grade tools. A free generative tool with no contract, no data commitments, and no admin controls has answered the Tier 4 questions by having no answers. The response is a rule your institution can enforce: no institutional or personal data goes into tools that have no data commitments. Module 5 turns that rule into the pre-deployment data checklist. What does not work is pretending the tools are not in the building; staff adopt them faster than procurement can see. Give staff an approved tool with real commitments, and the unapproved ones lose their audience.

Open source and self-hosted models. Where there is no vendor, the demands do not vanish, they change addressee: the 4.2 documentation demands are answered from the model's published documentation and by your integrator or host; the clauses that survive are the ones with the integrator or host; and the builder's obligations arrive with the model (Module 1's scope note: institutions that build or fine-tune hold both lists).

One rule that never scales down: a use case can outgrow its tier. A drafting aid adopted by the legal office, or a chatbot that starts answering benefits questions, has changed tier without changing product. Module 2, section 2.4 sets the re-tiering triggers; procurement's part is to record, for every purchase, the tier it was cleared at.

4.7 Where this module sits in existing standards

Specific article and clause references are collected, with verification, in the annex to this framework; the mapping at concept level:

This module	EU AI Act	NIST AI RMF	ISO/IEC 42001	OECD Principles / UNESCO Recommendation
4.2 documentation demands	Providers of high-risk systems must produce technical documentation and instructions for use; deployers are expected to operate the system according to them. Your tender demands are the buying-side mechanism for obtaining exactly this material.	The Map function: understanding context, capabilities, and limitations before use.	Expectations on information from suppliers and on documented system knowledge within the management system.	Transparency: made concrete as named documents rather than a value statement.
4.2.2 population evidence, subgroup breakouts	The Act's concern with accuracy and with discriminatory outcomes in high-risk uses.	The Measure function: performance evidence appropriate to context.	Impact assessment expectations.	Fairness and non-discrimination, turned into a question with a document behind it.
4.3 contract terms	Deployer obligations (use per instructions, monitoring, logging) are only performable if the contract secures the vendor cooperation each one needs.	The Govern function: third-party risk expectations.	Supplier and third-party controls within the management system.	Accountability: a named party owes a named obligation with a remedy.
4.4 and 4.5 vendor evaluation	Supports deployer duties of care in selection.	Govern and Map: third-party risk identification.	Supplier evaluation expectations.	Accountability and transparency in practice.

If your institution is ever asked, by a regulator, a ministry, an auditor, or a board, how its AI procurement aligns with recognised standards, the completed checklist, the interview record, and the signed clauses from this module are the evidence.

Artifact 4a: Vendor due diligence checklist

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0. This checklist is an operating tool and it is not legal advice.

How to score. For each item, enter: **0** (absent, refused, or answered only verbally), **1** (partial, or promised in writing but not yet delivered), **2** (delivered in a written document that will be attached to the contract). Items marked ★ are disqualifying items: a score of 0 on any ★ item fails the checklist for a Tier 2 purchase regardless of the total.

Pass thresholds.

Tier (from the Module 2 worksheet)	Threshold
Tier 2 (high stakes)	No ★ item at 0, and at least 80 percent of all available points across sections A to F
Tier 3 (operational)	No 0 on items A1, B1, D1, D3, F1, and on E3 where Section E applies, and at least 60 percent of available points on sections A, B, D, F
Tier 4 (low stakes)	Items D2, F1, F3 all at 2. Other sections optional

Section A: System documentation

#	Item	Score
A1 ★	System description document: intended use, excluded uses, inputs, outputs, system type	
A2	General training data description: time period, geography, languages, population coverage	
A3 ★	Written limitations statement: degradation conditions, failure modes, untested settings	
A4	Instructions for use suitable for training your oversight staff (Module 3)	

Section B: Performance evidence

#	Item	Score
B1 ★	Every performance claim in writing, each with metric, dataset, and population described	
B2 ★	Written answer to the population question: evidence of measurement on a population like yours, or an honest statement that none exists	
B3	Subgroup performance breakouts relevant to your setting, or a written statement of what was not measured	
B4	Method described in enough detail to be checked	
B5 ★	Written agreement to acceptance testing on your data, against criteria to be fixed in the contract	

Section C: Vendor interview (section 4.4)

#	Item	Score
C1	Interview held, answers recorded verbatim, record in the procurement file	
C2	Written follow-ups received for every question deferred in the interview	
C3	Vendor confirmed (question 16) that its answers will attach to the contract	
C4	Reference calls held with at least two comparable customers, notes on file	

Section D: Data, security, and logging

#	Item	Score
D1 ★	Written statement of storage, processing, and backup locations, by country	
D2 ★	Written answer on training use of your data and your users' inputs, with an off switch	
D3 ★	Audit logging confirmed in writing: what is logged, access, retention, export format	
D4	Security certifications provided with scope statements covering this product	
D5	Named subcontractor list with access and locations	
D6	Incident history for this product (three years) and notification commitment in hours or days	

Section E: Generative tools only (skip for purely predictive systems)

#	Item	Score
E1	Measured factual-error rate with method	
E2	Prompt injection testing evidence	
E3 ★	Training exclusion for your users' inputs, available and confirmed in writing	
E4	Per-user logging, exportable by your institution	

Section F: Contract posture

#	Item	Score
F1 ★	Exit and portability clause agreed: formats, days, fixed cost, survival	
F2 ★	Audit rights clause agreed	
F3 ★	Update and change control clause agreed: notice days, changelog, test window, version hold	
F4	Documentation-as-warranty clause agreed; entire-agreement wording checked accordingly	
F5	Breach notification hours and incident notification days written into the contract	
F6	Liability cap written as a currency figure; compared against the estimated cost of the worst realistic failure of this system; data breach liability confirmed as sitting outside or above the general cap; local counsel review completed and dated	
F7	Named termination triggers in the contract	

Red flag check (section 4.5). Number of red flags present: _____. One flag: purchase paused until resolved in writing. Two or more on a Tier 2 purchase: process ends with this vendor.

Result and sign-off

System and vendor	
Risk tier (attach the Module 2 worksheet)	
Total score / available	
★ items at 0	
Result (pass / fail / paused)	
Completed by (name, role, date)	
Approved at procurement gate (name, date)	

Artifact 4b: Contract clause library

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0.

Read this before using any clause. These clauses are drafting guidance written in plain language. They are not legal advice, and contract law differs by country. A clause that works in one jurisdiction may need different words, or may be unenforceable, in another. Before signature, have a lawyer qualified in your jurisdiction adapt and review the final text. Bring them this library and the vendor's draft together; the comparison will focus their work and reduce their fee. In these clauses, "the Institution" is your organisation, "the Vendor" is the seller, and "the System" is the AI system being purchased. Bracketed items in [square brackets] are numbers or lists you must fill in; a clause with an empty bracket is not finished. Each clause carries a marker: **(M)** mandatory, where walking away beats signing without them for a Tier 2 system; **(R)** recommended, concede only knowingly and in writing.

Clause 1: Documentation as warranty (M) The documents listed in Schedule [x] (including the system description, the limitations statement, the performance evidence, and the Vendor's written tender and interview answers) form part of this contract. The Vendor warrants that they are accurate and complete in all material respects. A material inaccuracy in any of them is a breach of this contract. *What this protects you from:* the sales claims evaporating at the moment you need them; an "entire agreement" clause quietly deleting everything you were told before signing.

Clause 2: Acceptance testing (M) The System will not enter live use, and the payment milestone in Schedule [x] will not fall due, until the Institution has tested the System on its own data against the acceptance criteria in Schedule [x] and the System has met them. The Vendor will provide the cooperation, configuration, and technical support needed to run the test. If the System fails, the Institution may require remediation and a retest, renegotiate, or terminate under Clause 11 with a refund of amounts paid for the System. The test will begin within [x] days of signature and complete within [x] days; if the Vendor's cooperation is not provided in time to meet those dates, the milestone does not fall due and the Institution may terminate under Clause 11. Acceptance or rejection is given in writing by the Institution's System Owner. Where fees are paid by subscription rather than milestones, the first [x] months' fees are refundable in full if the System fails acceptance. *What this protects you from:* discovering, as the buyers in section 4.1 did, that claimed performance and delivered performance are different systems.

Clause 3: Ongoing performance warranty (R) The Vendor warrants that, in live use, the System will perform within the ranges in Schedule [x], which states each metric by name (for example: accuracy, error rate by subgroup, alert or exception rate, response time), each with its measurement method. If measured performance falls outside the warranted range and is not restored within [x] days of notice, the Institution may terminate under Clause 11. *What this protects you from:* the system that passes on day one and degrades quietly in service, with no lever attached to the decline.

Clause 4: Audit rights (M) The Institution, or an independent auditor it appoints under confidentiality obligations, may audit the System's logs, the documentation, and the Vendor's compliance with this contract's data and security commitments: at least once per year, and additionally after any incident or suspected breach. The Vendor will provide access and cooperation, with audit logs and compliance documentation made available within [72] hours of a written request. The Institution gives [x] days' notice for routine audits and none for incident audits; the auditor is bound in writing to confidentiality; audits are conducted so as not to disrupt the service; and each party bears its own costs, except that audits beyond the first in any year are at the Institution's cost. *What this protects you from:* commitments that are true only in the sales meeting.

Clause 5: Update and change control (M) The Vendor will give the Institution at least [30] days' written notice of any change to the model or to System behaviour, with a written description of what is changing and why. The Institution may test the change before it is applied to its live environment and may decline it, remaining on its current version, for up to [x] months. Emergency security fixes may be applied immediately, with notice and description to follow within [x] days. If an applied change degrades the System against Schedule [x], the Vendor will restore the Institution's prior version (rollback) within [x] days of notice, at no charge. *What this protects you from:* silent updates: the validated system being replaced, unseen, by a system nobody validated.

Clause 6: Data ownership and permitted use (M) All data the Institution or its users provide to the System, and all outputs generated from it, remain the property of the Institution. The Vendor will use this data only to provide the contracted service. The Vendor will not use it to train, improve, or develop any model or product, and will not sell, share, or derive products from it, in original, aggregated, or anonymised form, without a separate written agreement signed by the Institution. *What this protects you from:* your institution's records, and your users' inputs, becoming someone else's asset by default. *Check locally:* who can own AI-generated output, and what rights exist in it, varies by jurisdiction; have counsel adapt the ownership words to your law. The clause's enforceable core is the restriction on the Vendor's use, which does not depend on ownership doctrine.

Clause 7: Data location and subcontractors (M for Tier 2) The Institution's data will be stored, processed, and backed up only in the countries listed in Schedule [x], and accessed only by the Vendor personnel and the subcontractors listed in Schedule [x]. Changes to either list require [x] days' advance written notice and, for the countries list and (for Tier 2 systems) the subcontractor list, the Institution's written consent. The Vendor will impose the obligations of this contract on every subcontractor and remains fully responsible for their performance. *What this protects you from:* your data migrating to jurisdictions, and into hands, you never evaluated; promises that stop at the vendor's front door.

Clause 8: Security and breach notification (M) The Vendor will maintain the security controls in Schedule [x] throughout the contract. The Vendor will notify the Institution of any security breach affecting the Institution's data within [x] hours of discovery, with a written account of scope, cause, and remediation to follow within [x] days. *What this protects you from:* learning about a breach after your own legal notification deadlines have already expired. Set the hours after checking your local breach notification law.

Clause 9: Product incident notification (R) If the Vendor becomes aware of a material defect, error, or safety problem in the System or in the same product deployed elsewhere, it will notify the Institution within [x] days, with its assessment of relevance to the Institution's deployment. *What this protects you from:* being the last customer to learn what the vendor already knows.

Clause 10: Audit logging (M for Tier 2) The System will log, at minimum: each input received, each output produced, the date and time, the model version, and the identity of the requesting user or process. Logs will be retained for [x] months, remain accessible to the Institution throughout, and be exportable by the Institution in the documented format in Schedule [x] at no additional charge. Logs will be protected against alteration, and any gap or edit in the log record is itself notifiable under Clause 8. *What this protects you from:* being unable to answer an affected person, an auditor, or your own incident investigation, because the record of your system's decisions belongs to someone else.

Clause 11: Termination triggers (R) In addition to any other rights, the Institution may terminate this contract if: the System fails acceptance testing under Clause 2; the System fails the performance warranty under Clause 3; the Vendor applies a change in breach of Clause 5; the Vendor refuses or obstructs an audit under Clause 4; the Vendor materially breaches Clauses 6, 7, or 8; or the Vendor announces withdrawal or end of support of the System or of the model version relied on. Termination under this clause entitles the Institution to a refund of fees paid for undelivered or unusable periods. *What this protects you from:* a contract where every exit requires proving "material breach" from a standing start, in a dispute, at your expense.

Clause 12: Exit and portability (M) On termination or expiry for any reason, the Vendor will: (a) export and deliver to the Institution all Institution data, configurations, and audit logs, in the open, documented formats in Schedule [x], within [x] days, for the fixed fee in Schedule [x]; (b) provide transition assistance for [x] months at the rates in Schedule [x]; and (c) then permanently delete all Institution data from its systems and its subcontractors' systems and certify the deletion in writing. The obligations in this clause survive termination and apply even while any dispute between the parties is unresolved. The export format will be validated during acceptance testing. Where the System is critical to continuing operations, the parties will record in Schedule [x] an escrow or continuity arrangement covering Vendor insolvency or withdrawal of the product. *What this protects you from:* lock-in, which is the quiet tax on every renewal; and a vendor holding your data hostage during a dispute.

Clause 13: Intellectual property indemnity (M for generative systems) The Vendor will defend and indemnify the Institution against third-party claims that the System, or output generated by the System when used as documented, infringes intellectual property rights, and will at its own cost secure the Institution's right to

continue use, modify the System to be non-infringing, or refund and assist exit under Clause 12. *What this protects you from:* your institution answering for what the vendor's training data contained.

Clause 14: Price protection and product continuity (R) The Vendor's charges for renewal terms will not increase by more than [x] percent per year, or by more than [named index], whichever is lower. The Vendor will give at least [x] months' written notice before withdrawing, discontinuing, or ending support of the System or of the model version relied on, and throughout that notice period exit assistance under Clause 12 remains available at the rates in Schedule [x]. *What this protects you from:* a dependency being repriced once switching has become expensive, and a product disappearing faster than you can leave it.

References for this module

- Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science*, volume 366, issue 6464 (25 October 2019), pages 447 to 453. doi:10.1126/science.aax2342.
- Habib, A. R., Lin, A. L., and Grant, R. W. "The Epic Sepsis Model Falls Short: The Importance of External Validation." *JAMA Internal Medicine*, volume 181, issue 8 (2021), pages 1040 to 1041. doi:10.1001/jamainternmed.2021.3333. (The editorial accompanying the Wong study; source of the developer-reported AUC range of 0.76 to 0.83.)
- Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestruie, J., Phillips, M., Konye, J., Penzoza, C., Ghous, M., and Singh, K. "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients." *JAMA Internal Medicine*, volume 181, issue 8 (2021), pages 1065 to 1070. doi:10.1001/jamainternmed.2021.2626.

Module 5: Data, security, and sovereignty

The moment your institution's data leaves the building for a vendor's system, control over it changes hands. This module gives you the questions to settle, and the checklist your Data Steward signs, before that moment.

Work it in order for each new system: draw the control line (5.1), inventory what would leave (5.2), settle where it goes (5.3), set the security requirements (5.4), cut and time-limit what you send (5.5), and check the special category notes if they apply (5.6). The output of all six sections is one signed document: the pre-deployment data checklist (Artifact 5), which blocks the readiness gate (G2, Module 3) until it is complete.

5.1 What you control, and what you hand over

In 2015, a London hospital trust agreed to provide patient records to a technology company for the clinical safety testing of a new application that would alert clinicians to acute kidney injury. Records of around 1.6 million patients were provided. In July 2017, the United Kingdom's data protection regulator ruled that the trust had failed to comply with data protection law: patients had not been adequately informed, no proper legal basis had been established, the volume of records provided was found excessive against the testing purpose, and no privacy impact assessment had been completed before the data moved. The trust was required to sign an undertaking committing it to change. (Information Commissioner's Office ruling on the Royal Free NHS Foundation Trust, 3 July 2017; references at the end of this module.)

The regulator's findings landed on the hospital. The technology company built the application, but the duty to inform patients, the duty to establish a legal basis, the duty to send no more than necessary, and the duty to assess the risks before transfer all belonged to the institution that held the data and decided to hand it over. Every question in this module exists because you will answer for the data after it leaves, so you must settle its terms before it leaves.

You control directly	You hand to the vendor
Which systems get which data, field by field	Physical storage and its locations
Who in your institution uses the system, and the acceptable-use note (Module 2)	Processing, backups, and infrastructure
The contract: every data, security, and deletion commitment (Module 4)	Which vendor staff and subcontractors touch the data
Configuration choices the product exposes: retention settings, training opt-outs, logging, region selection	Day-to-day execution of the promised security controls
Verification: audits, certificates, log exports (Module 4, Clause 4)	Incident detection on the vendor's side, until they notify you
The exit: export, deletion, certification (Module 4, Clause 12)	

The right column is governed only by two instruments: the contract that made the promises enforceable (Module 4 wrote it) and the verification that checks the promises are being kept (audit rights, certificates with scope statements, and log exports).

5.2 Inventory what would leave the building

Before any tender is answered or any pilot begins, the Data Steward writes down, on the checklist at the end of this module, what data the system would receive. Field by field, from the vendor's own integration documentation (demand it under Module 4.2.1), because "patient records" or "customer data" is a category, and categories hide the fields that later become the problem. At enterprise scale, field by field means field group by field group: inventory by the vendor's interface specification (the fields the integration sends, which is the inventory, never the whole legacy schema behind it), group fields that share a sensitivity class and purpose, and spend the manual scrutiny on the sensitive and regulated classes.

For each field, three entries:

1. Sensitivity class. Use these five, and where your national law defines its own classes, add or substitute them and record which law:

- **Public or internal operational data:** already published, or operational data with no person in it (equipment readings, aggregate statistics).
- **Personal data:** anything relating to an identifiable person, including identifiers like account numbers, addresses, and device identifiers. **Check locally:** your data protection law defines this term; use its definition.
- **Sensitive personal data:** the categories your law gives special protection, commonly health, ethnicity, religion, biometrics, sexual life, and criminal records, and often children's data. These fields carry the heaviest duties and the sector notes in 5.6.
- **Regulated data:** data whose handling a sector regulator prescribes (banking secrecy, patient records, tax records, student records). Name the regulator on the checklist.
- **Confidential institutional data:** unpublished decisions, security arrangements, commercial terms, anything whose disclosure would harm the institution itself.

2. The necessity test, asked field by field. *Does the system need this field to do its job?* Put the question to the vendor in writing and keep the answer (a strong vendor has a data specification that answers it already). The regulator's finding in 5.1 included this failure at the level of whole records: the volume provided was excessive against the purpose.

3. The lawful basis entry, as questions for local advice. This framework cannot tell you what makes processing lawful in your country. It can tell you what to write down and take to the person who can: What is our legal basis for using this data for this purpose? Does this purpose differ from the purpose the data was collected for, and does that difference require notice, consent, or something else here? Must the people in the data be informed, and how? Is a formal risk assessment (many laws name one, such as a data protection impact assessment) required before transfer? The 5.1 case is what it looks like when an institution assumes these answers instead of obtaining them: the legal basis it relied on did not hold. For Tier 2 systems and any sensitive personal data, these questions go to qualified local advice before the checklist is signed.

Unstructured sources. Generative deployments ingest documents, mailboxes, transcripts, and free-text notes, where sensitive data is buried in prose and a field inventory has nothing to grip. Inventory these by source rather than field: one row per source (the mailbox, the document store, the case file set), recording the sensitivity class of the worst data realistically inside it, the scrubbing or redaction control applied before ingestion and who verified that control on a sample, and the necessity answer for the source as a whole. A source that cannot be scrubbed to match its purpose is minimised the other way: excluded.

5.3 Settle where the data goes, and whose law follows it

Residency means the countries where your data is physically stored, processed, and backed up. It matters for three reasons: your law may restrict where this data may go (many countries restrict cross-border transfer of personal, health, financial, or government data); the vendor's location decides which foreign laws and authorities can reach your data; and distance weakens every practical remedy you have when something goes wrong.

The questions to settle in writing, all of them demanded at tender time under Module 4.2.5 and locked by Clause 7 of the clause library:

1. In which countries is our data stored, processed, and backed up? (Three separate answers; backups are the one vendors forget to mention.)
2. From which countries can vendor and subcontractor staff access it? Support access is access: an engineer who can view your data to fix a problem is a data location, whatever the storage map says.
3. Which country's law governs the vendor, and can an authority in that country compel disclosure of our data? You are not expected to answer this yourself; you are expected to put it, in writing, to the vendor

and to your local adviser, and to file both answers. Require the vendor's written attestation of which foreign laws can compel disclosure of your data and how the vendor responds to such demands.

4. Does our law, or our regulator, restrict this data leaving the country or the sector? **Check locally**, and name the instrument on the checklist. Where the answer is yes, the conversation changes from "which region" to "which deployment model": many products offer in-country hosting or on-premises deployment (running on your own infrastructure), and if this vendor does not, that is a finding for the due diligence checklist.
5. If we later need the data moved (a region change, a law change, an exit), what does the contract say it costs and how long it takes? (Clause 12 of the Module 4 library fixes the export days and fee; a migration between vendor regions needs its own line in the contract schedule, so add one.)

One configuration rule that costs nothing: where the product offers a region selection, the Data Steward sets it, documents it, and the setting goes into the contract schedule rather than remaining a dashboard preference an update can silently reset.

5.4 Security controls that matter for AI systems specifically

Your institution's baseline IT security (access management, patching, backups, network controls) is assumed here; ISO/IEC 27001 is the recognised reference for that baseline and Module 4.2.5 already demanded the vendor's certificates with their scope statements. This section adds the controls that are specific to AI systems, each as a requirement you can write and a verification you can perform.

1. Access control on the system, its configuration, and its logs. Require: named-account access only (no shared logins) for anyone who can query the system, change its configuration, or read its logs; access reviewed quarterly by the Operational Lead; administrator actions logged. Verify: ask for the access list at go-live and at each quarterly review; the list should be short and every name explainable. The same discipline covers the system's secrets: API keys (the credentials software uses to call a service), service accounts, and the prompt or system instructions that configure a generative system's behaviour are access-controlled, never shared in documents or tickets, and rotated when a holder leaves.

2. Input attacks on systems that read text. A **prompt injection**, defined in Module 4.2.8, is an attack where instructions hidden in text the system reads (an email, an uploaded document, a web page) hijack its behaviour. Any generative or text-reading system that touches untrusted content (content from outside the institution, or content a member of the public can influence) is exposed by design. Require: the vendor's injection testing evidence (Module 4 question 15) and a written statement of the mitigations in the product. Verify: for public-facing deployments, commission an independent security test before go-live; this is the first of the two named places where you hire a specialist, and Module 1.4 already put it on leadership's cost list.

3. Leakage of what the system was fed. Two channels, one question each. Training leakage: can the system reveal, in its outputs, data it was trained or fine-tuned on? If your data is used for training or fine-tuning, your data can surface in outputs, potentially to other users; this is why Module 4's Clause 6 makes training on your data a separate written decision rather than a default. Session leakage: is one user's input visible to another user, another department, or another customer of the vendor? Require written answers to both, and for any multi-customer product, the vendor's written description of how customers are separated. Verify: at validation (Module 6), test with planted, harmless canary records (fabricated entries unique to your institution) and check they cannot be retrieved from another account or session.

4. The vendor's own access, bounded. Require in the contract: vendor staff access to your data only for named support purposes, logged, and reviewable by you under the audit clause; remote support sessions initiated or approved by your side for systems holding sensitive personal data. Verify: request the vendor access log at each annual audit and read it beside the support tickets; access with no matching ticket is a finding.

5. Query and output logging as a security instrument. Module 4 demanded audit logs for governance. They are also your intrusion evidence: who queried what, when, in what volume. Require: logs exportable to your side, and a monthly look (Operational Lead) at query volumes by account.

6. Manipulation of decision systems. Predictive systems that allocate scrutiny (fraud flags, inspection targeting, queue prioritisation) invite a specific behaviour: the people being scored learn the inputs and shape them. Require from the vendor a written statement of which inputs most influence outputs and what monitoring would reveal gaming of them; feed it into the Module 6 monitoring plan. For systems where successful manipulation could affect safety, security screening, or access to essential services (rather than costing money alone), add independent adversarial testing to the pre-deployment plan: the second of the two named places where a specialist is hired.

5.5 Send less, and put an end date on everything

Minimisation is a security control: data that never left cannot leak, cannot be subpoenaed abroad, cannot be retained past its welcome, and cannot appear in a regulator's findings. The 5.2 inventory's necessity test is the instrument; the rule of use is: send the minimum field set the vendor's own data specification justifies, and record what was withheld and why on the checklist.

Anonymisation. Two words that are treated as interchangeable and are legally and practically different.

Pseudonymisation replaces direct identifiers (names, account numbers) with codes; the person can be re-identified by whoever holds the key, and most laws treat pseudonymised data as still personal. **Anonymisation** is supposed to make re-identification impossible, and the state of the art is that rich individual-level records (dates, locations, diagnoses, transactions) are very hard to anonymise dependably, because combinations of ordinary attributes single people out. The operating rules that follow: treat "we anonymise your data" in a vendor document as a claim requiring the method in writing, and take that method to your local adviser for any sensitive personal data; prefer aggregation (sending counts and statistics rather than records) where the use case allows it, since well-aggregated data leaves individuals behind; and never let the word "anonymised" relax the residency, contract, or deletion requirements of this module on its own.

Derived representations are the data. Embeddings (numerical representations of text or records that retrieval systems search over), vector stores, caches, and a system's conversation memory are made from your data and can reveal it. For every question in this module (location, access, retention, deletion, training use), treat them as the data they were made from: a deletion certificate that does not cover the vector store has not deleted; a retrieval pipeline that sends passages of your documents to a vendor's model is a data flow to inventory like any other; and the region setting, the training exclusion, and the exit export all apply to the derived stores by name.

Retention. For each data flow on the inventory: how long does the vendor keep it, what happens at the end (deletion or return), and who confirms? The numbers go in the contract schedule, and the deletion certificate at exit is signed into the Data Steward's role (Module 3). A retention entry reading "for the duration of the service" is an intention; "deleted within 30 days of contract end, certified in writing" is a date, and the checklist accepts only dates. Outputs are data too: where a generated report, letter, or assessment enters a case file, it becomes a record under your retention rules, and the inventory records the flows outward the same as the flows in.

5.6 Special categories

Each note below names the boundary where this framework's general method stops being sufficient and a specialist or regulator enters. In every case the Module 5 checklist still applies in full; these notes add to it.

Health data. Nearly every jurisdiction gives patient data its own regime, and many require specific approvals, impact assessments, or ethics review before it is shared for anything beyond direct care. The 5.1 case sits here: testing and improvement are not direct care, and the distinction is where that institution's reasoning failed. Boundary: any flow of patient data to a vendor for testing, training, or product improvement goes to your health data regulator's guidance and to local advice before the checklist is signed.

Children's data. Most laws treat it as requiring special protection, with specific consent and age assurance rules, and several regulate systems directed at children as a category. Boundary: any system whose users or subjects are children (schools above all) gets the lawful basis questions in 5.2 answered by qualified local advice, without exception, and the vendor's experience with children's data becomes a due diligence line.

Financial data. Banking secrecy, payment card standards, and financial regulators' outsourcing rules commonly constrain where account and transaction data may go and impose notification or approval duties when critical functions are outsourced. Boundary: if a financial regulator supervises your institution, its outsourcing and cloud guidance governs this module's questions, and the checklist records the mapping; where the two conflict, the regulator wins.

Government administrative data. Tax records, benefits records, civil registries, and case files often carry statutory confidentiality that predates any data protection law, and their movement across borders can raise sovereignty questions no vendor assurance settles. Boundary: for these datasets, the residency questions in 5.3 are answered by your legal framework first and the vendor's architecture second, and in-country or on-premises deployment is the default position from which departures must be justified in writing.

5.7 Where this module sits in existing standards

This module	Where it maps
The inventory, necessity test, and lawful basis questions (5.2)	Data protection law's principles of lawfulness, purpose limitation, and data minimisation, wherever a law of that family applies to you; the EU AI Act's data governance expectations for high-risk systems; ISO/IEC 42001's data management and impact assessment expectations.
Residency and transfer questions (5.3)	National transfer restriction regimes (check locally, always); the contract mechanics live in Module 4 Clause 7.
AI-specific security controls (5.4)	The NIST AI RMF Measure and Manage functions applied to security risk; ISO/IEC 27001 as the assumed baseline for general security management, with this section as the AI-specific layer above it.
Minimisation, retention, anonymisation (5.5)	Data protection minimisation and storage limitation principles; the OECD Principles and UNESCO Recommendation on privacy and safety, made operational as field lists and dates.
Special category boundaries (5.6)	Sector regimes (health, children, financial, government), which prevail over this framework wherever they apply.

Artifact 5: Pre-deployment data checklist

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0. Completed and signed by the Data Steward before the readiness gate (G2, Module 3); an incomplete checklist blocks the gate. It is an operating tool and it is not legal advice; items marked “local advice” require answers from a person qualified in your jurisdiction, and their name and date go on the line.

System

System and vendor	
Risk tier (attach the Module 2 worksheet)	
Vendor data specification received (Module 4.2.1)	☐ attached

Section A: Inventory. One row per field or field group leaving the institution. Attach continuation sheets as needed.

Field	Sensitivity class (5.2)	Needed for function? (vendor's written answer)	Withheld instead?	Regulator, if regulated data

Section A2: Unstructured sources (generative deployments; one row per source)

Source	Worst sensitivity class realistically inside	Scrubbing or redaction control, verified on a sample by	Needed for function?	Derived stores (embeddings, caches, memory) covered by location, retention, deletion answers?

Section B: Lawful basis (local advice for Tier 2 and any sensitive personal data)

- ☐ Legal basis for this use identified and recorded: _____
- ☐ Purpose compared with collection purpose; notice or consent steps identified: _____
- ☐ People in the data informed, where required; how: _____
- ☐ Formal risk or impact assessment required? ☐ No ☐ Yes → completed and attached
 - Local advice by (name, qualification, date): _____

Section C: Location and access

- ☐ Storage countries, in the contract schedule: _____
- ☐ Processing countries: _____ Backup countries: _____
- ☐ Countries from which vendor or subcontractor staff can access: _____
- ☐ Transfer restrictions checked locally; instrument named: _____
- ☐ Region setting configured and recorded, where the product offers one
- ☐ Governing law of vendor and compelled-disclosure question put in writing to vendor and adviser; answers filed

Section D: AI-specific security (5.4)

- ☐ Named-account access only; access list at go-live filed; quarterly review scheduled
- ☐ Injection testing evidence received (systems reading untrusted text); independent security test commissioned for public-facing systems (name, date): _____

- ☐ Training use of our data: excluded in writing, or separately agreed with leadership signature attached
- ☐ Session and customer separation described by vendor in writing; canary test planned for validation
- ☐ Vendor access bounded in contract; access log review scheduled into annual audit
- ☐ Log export confirmed; monthly query-volume review by the Operational Lead
- ☐ Decision systems: influential-inputs statement received and gaming metric added to the monitoring plan

Section E: Minimisation and retention

- ☐ Field set reduced to the vendor specification's justified minimum; withholdings recorded in Section A
- ☐ Any "anonymised" or "aggregated" claim: method in writing, taken to local advice where sensitive
- ☐ Retention, as dates, per flow, in the contract schedule: _____
- ☐ Deletion or return at end, with written certification, per Module 4 Clause 12

Section F: Special categories (5.6), where applicable

- ☐ Health data: regulator guidance identified; approvals or ethics review obtained where required
- ☐ Children's data: lawful basis answered by local advice; vendor experience recorded in due diligence
- ☐ Financial data: supervisor's outsourcing and cloud rules mapped; notifications or approvals done
- ☐ Government administrative data: statutory confidentiality checked; in-country or on-premises default applied or departure justified in writing

Sign-off

	Name	Role	Date
Completed and signed		Data Steward	
Local advice items confirmed (where applicable)			

File with the G2 gate record in the system register (Module 7).

References for this module

- Information Commissioner's Office (United Kingdom), "Royal Free - Google DeepMind trial failed to comply with data protection law," news release, 3 July 2017, with the accompanying Data Protection Act 1998 undertaking signed by the Royal Free NHS Foundation Trust (ico.org.uk, undertaking document royal-free-undertaking-03072017). Findings drawn on: approximately 1.6 million partial patient records processed for the clinical safety testing of the Streams application; patients not adequately informed; a proper legal basis to be established; the volume provided found excessive; a privacy impact assessment required, including specific steps for transparency.
- ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection, Information security management systems, Requirements. Referenced as the assumed baseline standard in section 5.4.

Module 6: The deployment lifecycle

An AI system that passed every test on the day it went live will drift, be updated under you, and degrade while attention moves elsewhere. This module gives you the stage-by-stage controls, the clocks, and the thresholds that keep a running system as governed as a new one.

For a new system, work the stages in order: validate (6.2), roll out in phases (6.3), then run the standing machinery of monitoring (6.4), appeals (6.5), incidents (6.6), updates (6.7), and revalidation (6.8), until the day you retire it (6.9). For a system already live with none of this around it, start at 6.4 and 6.5 this month, then run 6.2 as a late validation at the next revalidation date. Artifact 6a is the stage checklist with the gates and durations; Artifact 6b is the monitoring plan; Artifact 6c is the notice that tells affected people how to challenge a decision.

6.1 The life of a system

A sepsis prediction model was deployed at hundreds of hospitals on the strength of developer-reported accuracy of 0.76 to 0.83. In 2021, researchers tested it on their own hospital's 38,455 admissions and measured 0.63, with two thirds of sepsis cases missed and alerts on nearly one patient in five (Wong and colleagues, cited in full in Module 4). In October 2022, reporting based on the developer's own documents showed the model had been overhauled, and the developer was now recommending something it had not required before: that the model be trained on each hospital's own data before clinical use. A revised version released that year was built to allow that local fine-tuning. (References at the end of this module.)

The gap between claimed and delivered performance was findable by validation on local data before go-live (6.2). The flood of alerts that made oversight impossible was measurable in the first weeks of a shadow run (6.3) and in monitoring afterwards (6.4). The missed patients were the failures that produce no alert, findable only by deliberately auditing what the system stayed silent about (6.4). And the overhaul is the reminder that sits under section 6.7: the system you validated is a version, vendors replace versions, and an update arriving is a normal, certain event.

The lifecycle at a glance. Artifact 6a is this table as a working checklist.

Stage	What happens	Ends at	Typical duration (Tier 2 defaults)
Validation (6.2)	The system is tested on your own historical data against criteria fixed in advance	G2 readiness (with the Module 5 data checklist)	2 to 6 weeks
Shadow mode (6.3)	The system runs live; humans decide as before; outputs are compared to reality	Go-live decision input	Minimum 4 weeks and at least 200 outputs; 6 to 8 weeks where detectability is low
Pilot (6.3)	Limited live use, pre-agreed scope and stop criteria	G3 go-live for full scope	4 to 12 weeks
Operation (6.4 to 6.7)	Monitoring, appeals, incidents, update control	G4 continued operation, every 12 months	Ongoing
Revalidation (6.8)	The 6.2 test re-run on fresh data	G4	1 to 2 weeks, annually or on trigger
Decommissioning (6.9)	Ordered shutdown with data return	Register closure	2 to 8 weeks

Elapsed time from contract signature to full go-live for a Tier 2 system, under these defaults: typically three to six months. Publish that expectation to leadership at G1. Small institutions: where 300 cases or a two-stage rollout exceeds what you have, the path shrinks: state on the results what a smaller sample cannot show, run shadow mode longer in place of a separate pilot, and keep every gate.

6.2 Validation

What validation is here: running the system against a sample of your institution's own historical cases, where you already know what the right answer was, and comparing its performance to criteria you wrote down before you saw any results. No data science team is required. The method:

1. Fix the criteria first, in writing, at contract time. They live in the contract's acceptance schedule (Module 4, Clause 2), so the consequence of failure is already agreed. Three kinds of criterion:

- How often the system must be right (overall, and for the outcome that matters most, such as the cases that should be flagged).
- The worst subgroup gap you will accept (performance broken out by the groups chosen under the Module 4.2.2 rule: what you already report on, plus what your law protects, defaulting to age, sex, and language).
- The operational limits (an alert or exception rate the receiving team can absorb, which the sepsis arc above shows is a safety property rather than a convenience).

2. Assemble the test sample. From your records: at least 300 cases for a Tier 2 system, and at least 30 in each subgroup you intend to report on (framework defaults; below those numbers, say on the results what the sample cannot show), drawn to resemble the population the system will serve (all sites, all seasons or cycles if your work is seasonal, and enough members of each subgroup in step 1 to say something about them; where a subgroup is small, say so in the results rather than silently dropping it). Include known hard cases and past errors on purpose.

3. Freeze the sample and keep it. This set is an instrument, and you will use it again at every vendor update (6.7) and every revalidation (6.8). Store it under the Data Steward's control, and never let it flow into any vendor system as training or improvement data, or the instrument is contaminated and measures nothing.

4. Run the system on the sample, with the vendor's cooperation the contract already promised, in a test configuration that cannot touch live operations.

5. Score against the criteria from step 1, and only those. The discipline of deciding the pass mark before seeing the results is the integrity of the exercise.

6. Compare against current practice. Is the system better than what we do today, on these same cases? A tool that matches your current error rate while adding a vendor, a data flow, and a monthly governance load has a negative case, in writing.

7. Record, decide, and file at G2. Pass: proceed to shadow mode. Fail: the contract's acceptance clause already names your options, so use them in order:

- Require remediation and retest.
- Renegotiate scope to the uses the results support.
- Exit with your money.

What must never happen is the quiet fourth option, going live anyway with a note to watch it closely.

Where no ground truth exists. Some systems produce outputs that reality never grades (plans, allocations, drafts, designs). Score the system's outputs blind against your current process's outputs on the same cases, judged by a small panel of your own domain experts against criteria fixed in advance, and let shadow mode (6.3) carry more of the weight. State in the results which comparator was used.

6.3 Phased rollout

Shadow mode: the system runs on live cases in real time, its outputs are recorded, and nobody acts on them; humans decide exactly as they did before, and at the end of each period the system's outputs are compared against what the humans did and, where outcomes have landed, against what turned out to be true. It is the only stage that measures the system against reality at zero risk to any person, and it is the default first live stage for every Tier 2 system, mandatory where the Module 2 worksheet marked detectability low.

Run it for at least 4 weeks and at least 200 outputs, whichever takes longer (6 to 8 weeks where detectability is low), and use it to capture three numbers you can get nowhere else:

- The live alert or exception rate on your real caseload (the operational criterion from 6.2 tested against reality).
- The minutes-per-case cost of real review, which Module 3.4 turns into the reviewer time budget.
- The first version of the monitoring baselines (6.4).

Shadow results go to the same criteria discipline as validation: thresholds first, results second.

Pilot: limited live use, defined in one written paragraph before it starts: the scope (one site, one team, one case type), the duration (4 to 12 weeks), the success criteria, and the stop criteria, with the stop criteria carrying named numbers (“pause the pilot if the override rate exceeds x percent, if any Severity 1 incident occurs, or if weekly volume exceeds y”). The pilot ends at a decision, recorded either way.

Scale: expansion beyond the pilot’s written scope is a G3 gate signature (System Owner and Independent Reviewer). Each further major expansion (a new site, a new population, a new decision type) repeats in miniature: check the Module 2 worksheet still describes the use, then a short shadow or pilot period sized to the change.

6.4 Monitoring

Monitoring is a one-page plan per system (Artifact 6b), written before go-live, with columns: what is measured, how often, who looks, and the two thresholds at which someone must act. **A monitoring report nobody is named to read is not monitoring**, which is why the plan’s “who looks” column carries names from the Module 3 RACI rather than team titles.

What to measure. The default metric set for a Tier 2 system; delete what does not apply and say why on the plan:

1. **Volume:** outputs per week, by site or channel.
2. **Output distribution:** the rates of each outcome class (flagged and unflagged, approved and referred, each score band). **Drift** is the slow movement of these rates away from the baselines captured in shadow mode, without anyone having changed anything; it happens because the world the system was trained on stops matching the world it runs in.
3. **Human interaction rates:** overrides, exceptions, and written-reason patterns, per Module 3.5, with its two investigation thresholds (below 1 percent and above 30 percent) copied onto this plan.
4. **The failures that produce no alert, counted deliberately.** A system’s visible errors announce themselves; its silent ones (the patient never flagged, the claim never selected, the fraud never scored) are invisible because the system’s silence looks the same whether it is right or wrong. Three mechanisms count them:
 - The monthly unflagged-case audit (pull a sample of cases the system did not act on, 20 cases or 2 percent monthly, whichever is larger, and have a human check what happened to them).
 - The blind second review sample already running under Module 3.5.
 - Outcome follow-up where reality eventually reports the answer (the missed sepsis case surfaces in the clinical record; the missed fraud surfaces in later audits), which means someone is named to reconcile outcomes against the system’s old outputs quarterly.
5. **Subgroup performance,** on the same groups as validation, wherever outcome data allows; where it does not, the unflagged-case audit sample is stratified by those groups instead, so the question is still being asked.
6. **Complaints and appeals** (6.5): counts, themes, and outcomes, as a monitoring input on equal footing with the technical metrics, because the affected person is often the only sensor that ever meets the error.
7. **Input data quality:** missing fields, malformed records, and source changes. (The Data Steward’s change-approval duty in Module 3 is the detection mechanism for source changes: any approved change to sources appears in this row’s notes for the month it happened.)

Cadence and eyes. Operational Lead: weekly glance at volume and interaction rates for low-detectability systems, monthly full review for every Tier 2 system, producing the one-page summary the System Owner receives (Module 3). Tier 3: monthly against actuals. Tier 4: an annual sanity check.

Thresholds, two levels, written before go-live. For each metric: an **investigate** level (someone looks within 5 working days and records what they found) and an **act** level (the Operational Lead's pause authority is triggered and the incident question is asked the same day). The plan template forces both numbers at go-live.

6.5 Complaints and appeals from affected persons

As a matter of fairness, someone who loses care, money, benefits, or standing to a decision a system shaped is owed a route to a human who can look again; the values instruments this framework maps to (Module 1) all say so. And as a matter of engineering, section 6.4 already counted the appeal channel among the monitoring instruments.

Design the channel before go-live; the G3 gate checks it exists. Five design points, each with its default:

1. The notice. Wherever a person meets a decision the system informed (the letter, the portal message, the counter), they are told, in plain language: that they can ask for the decision to be looked at again, how (a named channel: a form, an address, a phone line, a counter), and by when they will hear back. Artifact 6c is the template. For conversational systems there is no decision letter, so the channel is built into the interface instead: a visible route to challenge an answer on every session, carrying the session identifier as the reference, with the same clocks, and the notice text published where the chat is offered. **Check locally:** some laws require telling people that an automated system was involved in a decision, and some grant explicit rights to human review; your notice must satisfy whichever applies, and this framework's notice is the floor.

2. The clock. Acknowledgement within 5 working days; a substantive answer within 20 working days (framework defaults). Where your law or sector rules set shorter deadlines, those govern; write the governing number on the notice and hold it.

3. Fresh eyes with real information. The appeal is decided by a person who did not make or review the original decision, working with the Module 3.4 information set: the inputs, the system's output and whatever it can show about its reasoning, the original reviewer's recorded reason, and the underlying record. The reviewer decides the case on its merits, and their authority to depart from the system's output is the same authority Module 3.4 defined, with the same protection.

4. The log. Every appeal is logged: received, decided, outcome (upheld, varied, or confirmed), and the one-line reason. The log feeds the monitoring plan, and themes get read monthly by the Operational Lead beside the technical metrics.

5. The trigger. An upheld appeal that reveals a systematic error (the same failure would affect other people) is an incident under 6.6 the day it is upheld. Three upheld appeals traceable to the system in a quarter (a framework default) trigger the incident question automatically, whatever the individual explanations.

The cost of all five points, for most institutions, is a template, a mailbox, a log, and a duty rota measured in hours per month.

6.6 Incident response

What counts as an incident. Any of the following, and the list is deliberately closed:

1. A wrong or harmful output reached an affected person, including through an oversight bypass (the Module 3.6 bypass rule points here).
2. A systematic error is discovered by any route: validation, monitoring, the unflagged-case audit, an upheld appeal under 6.5, or staff report.
3. A security event touches the system or its data (coordinate with the institution's general security process; the Module 5 controls are the evidence sources).

4. The vendor notifies a material defect, its own breach, or an incident at another customer of the same product (the Module 4 Clauses 8 and 9 made them tell you).
5. An unnotified change to the model or to system behaviour is discovered (see 6.7). Routine maintenance of the surrounding software that changes neither the model nor its behaviour is not an incident; it is the short path in 6.7 step 2.

Severity, aligned with the contract's definitions (Module 4.3.12): Severity 1, wrong or harmful outputs are reaching people now; Severity 2, the system is unavailable or known to be producing unreliable outputs but they are contained; Severity 3, degraded, suspected, or near-miss.

The response steps, with clocks.

Step	What	Who	Clock (Severity 1 / 2 / 3)
Contain	Pause the system or the affected output channel; stop the harm first and investigate second	Operational Lead (48-hour pause authority), any reviewer for a single output (Module 3.6)	Immediately / same day / at next review
Declare	Named incident record opened from the Module 7 template	Operational Lead declares; System Owner informed	Within 4 working hours / 24 hours / 5 working days
Assess	What happened, who was affected, since when; the audit logs demanded in Module 4.2.4 are the evidence	Operational Lead, with the vendor's cooperation under contract	First written assessment within 5 working days
Notify	Internal: System Owner and leadership per severity. External: affected persons, and regulators where required. Check locally: breach and incident notification duties and deadlines vary; your Module 5 checklist named the instruments, and the vendor's notification clock (Module 4, Clause 8) was set to run faster than yours for exactly this moment	System Owner owns the decision; Data Steward for data-protection notifications	Per the governing law's clock, decided within the assessment window
Remediate	Fix, retrain, reconfigure, or restrict; where people were affected, correct the decisions themselves: where law permits, reopen them proactively through the appeal machinery in 6.5 rather than waiting to be asked, and where the decisions carry legal finality (settled assessments, concluded proceedings, closed accounts), take the reopening question to counsel first, because in some sectors unilateral reopening is itself unlawful	System Owner directs	Plan within 10 working days
Record and learn	Close the incident with a written record: cause, effect, changes made; file in the register; feed any new failure mode into reviewer training (Module 3.4) and the monitoring plan	Operational Lead	At closure

Restart. A paused Tier 2 system restarts only on the System Owner's signature, after the relevant slice of validation (6.2) or the frozen test set (6.7) has been re-run and passed.

6.7 Vendor updates

The arc in 6.1 ends with the vendor itself overhauling the model and moving to local training: versions change, materially, under every deployer. Module 4's Clause 5 bought you the rights this section spends: notice, a changelog, a test window, and the ability to stay on your version. The operating routine, run by the Operational Lead for every update notice:

1. Log it. Every update notice goes in the system's change log with its date, the vendor's changelog, and the planned decision date inside the notice window.

2. Read the changelog against your use. Three questions, recorded in one paragraph: Does the change touch the model or only the surrounding software? Does it touch the inputs, outputs, or decision logic of our use case? Does the vendor's own description mention changed behaviour, retraining, or new data? Software-only changes with no behavioural claims can take the short path (apply in a maintenance window, watch the monitoring for two weeks). Everything else takes the full path. Vendors on continuous deployment pipelines push software changes constantly; the contract's notice clause governs model and behaviour changes, and this routine is scoped to match.

3. Re-run the frozen test set. The validation sample frozen in 6.2 step 3 is re-run against the new version in the test window, and scored against the same criteria as the original acceptance. This is a day or two of work. For generative tools, the equivalent instrument is a **frozen prompt set**: 30 to 50 representative tasks from your actual use, each with notes on what an acceptable answer contains, assembled during validation and re-run on every model change; generative updates change behaviour broadly.

4. Decide, in writing, inside the window.

- Accept (results hold: apply, update the register with the new version identifier, brief reviewers on any changed failure modes per Module 3.4).
- Defer (results worsen or are unclear: stay on the current version under the contract's version-hold right, tell the vendor what failed, retest their fix).
- Refuse and escalate (the vendor cannot or will not fix, and the version-hold period is finite: this goes to the System Owner as a G4-level decision, because a permanently unacceptable update is the beginning of decommissioning, 6.9).

5. Treat silent updates as incidents. If monitoring moves and the investigation finds the model changed without notice, that is incident type 5 under 6.6 and a breach of Clause 5: contain, document, and put the vendor's explanation in writing.

Emergency security patches take the contract's exception lane: applied fast, with notice following, and the frozen test set run immediately after application rather than before.

6.8 Periodic revalidation

Once a year for Tier 2 (24 months for Tier 3), and immediately on any of these triggers:

- A Module 2 re-tier trigger fires.
- Monitoring crosses an act threshold.
- An accepted major update (the 6.7 re-run counts as the revalidation if full).
- The served population or upstream data sources change materially.
- An incident closes with cause unknown.

The exercise is 6.2 in light form: refresh the frozen sample with recent cases (retire the oldest, keep the hard ones, keep the subgroup coverage), re-run, score against the same criteria, and file at G4. One to two weeks of elapsed time. A revalidation that fails goes down the same three-option path as a failed acceptance, with the contract's performance warranty (Clause 3) now doing the work Clause 2 did at the start.

6.9 Decommissioning

The System Owner takes the decommissioning decision, in writing, when any of these occurs:

- A failed revalidation the vendor cannot cure.
- An update impasse under 6.7.
- The contract's exit triggers (Module 4, Clause 11).
- A better alternative passing its own validation.
- The purpose itself ending.

From decision to closure, the sequence, in order:

1. **Fallback first.** The manual or predecessor process has handled at least one full day's real caseload, or a realistic rehearsal of one, within the month before switch-off, and the staffing for it is named.
2. **People informed:** users and operators, with dates; affected persons where the change alters how decisions about them are made (the 6.5 notice channel carries it).
3. **Data back:** the full export under Clause 12 (data, configurations, audit logs) received and verified readable on your side.
4. **Decisions preserved:** the record of what the system decided, and the logs behind it, are retained under your institution's retention rules, because appeals, audits, and legal questions outlive the system that caused them. **Check locally** for the governing retention periods.
5. **Deletion certified:** the vendor's written deletion certificate, covering subcontractors, signed into the file by the Data Steward.
6. **Access closed:** accounts, integrations, and data feeds shut off, on a list, each item initialled.
7. **Register closed:** the system's entry marked decommissioned, with dates and the reason.
8. **One page of lessons:** what the institution would demand, test, or write into the contract differently next time, filed where the next procurement will find it: with the Module 4 checklist templates.

6.10 Where this module sits in existing standards

This module	Where it maps
Validation and phased rollout (6.2, 6.3)	The EU AI Act's accuracy and testing expectations for high-risk systems, and the deployer's duty to use systems per their instructions; the NIST AI RMF Measure function (test and evaluate before and during use).
Monitoring and drift (6.4)	The Act's deployer duty to monitor operation; NIST AI RMF Measure and Manage; ISO/IEC 42001's operation and performance evaluation clauses.
Appeals (6.5)	The transparency and redress themes of the OECD AI Principles and the UNESCO Recommendation; national rights to human review of automated decisions, where they apply (check locally; the annex collects the references).
Incidents (6.6)	The Act's serious incident reporting expectations for the AI ecosystem; national breach and incident notification regimes (check locally).
Update control and revalidation (6.7, 6.8)	ISO/IEC 42001's change management and continual improvement expectations; the AI RMF Manage function.
Decommissioning (6.9)	ISO/IEC 42001's lifecycle view of the management system; the contract mechanics live in Module 4.

Artifact 6a: Stage-by-stage lifecycle checklist

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0. One per system; items initialised as completed; gate rows signed per Module 3. Durations are Tier 2 defaults.

Stage 1: Validation (2 to 6 weeks)

- ☐ Acceptance criteria fixed in the contract schedule before testing (accuracy, worst subgroup gap, operational limits)
- ☐ Test sample assembled: at least 300 cases, representative, subgroups covered, hard cases included
- ☐ Sample frozen under Data Steward control; excluded in writing from vendor training
- ☐ Run completed in a test configuration; results scored against pre-fixed criteria only
- ☐ Compared against current practice on the same cases
- ☐ Decision recorded: pass / remediate and retest / renegotiate scope / exit
- ☐ **GATE G2:** Data checklist (Module 5) and validation results signed: Data Steward _____ Independent Reviewer _____ Date _____

Stage 2: Shadow mode (minimum 4 weeks and 200 outputs; 6 to 8 weeks if low detectability)

- ☐ Outputs recorded, nobody acting on them; comparison method defined first
- ☐ Live alert or exception rate measured against the operational criterion
- ☐ Real minutes-per-case review cost measured; reviewer time budget set (Module 3.4)
- ☐ Monitoring baselines captured for every Artifact 6b metric

Stage 3: Pilot (4 to 12 weeks)

- ☐ One-paragraph pilot definition written first: scope, duration, success criteria, stop criteria with numbers
- ☐ Oversight design one-pager in force (Module 3.4); reviewers trained on limitations and validation failures
- ☐ Appeal channel live and notice published (Artifact 6c)
- ☐ Incident plan and monitoring plan (Artifact 6b) in force
- ☐ Pilot decision recorded either way
- ☐ **GATE G3 (full go-live):** System Owner _____ Independent Reviewer _____ Date _____

Stage 4: Operation (standing)

- ☐ Monthly monitoring review and summary to System Owner; weekly glance if low detectability
- ☐ Monthly unflagged-case audit (20 cases or 2 percent); quarterly outcome reconciliation
- ☐ Appeals log read monthly beside the metrics
- ☐ Every update notice through the 6.7 routine; change log current; frozen test set or prompt set re-run on model changes
- ☐ **GATE G4 (every 12 months and on trigger):** revalidation passed, roles filled, thresholds reviewed: System Owner _____ Date _____

Stage 5: Decommissioning (2 to 8 weeks)

- ☐ Trigger and decision recorded
- ☐ Fallback confirmed working and staffed before switch-off
- ☐ Users, operators, and (where relevant) affected persons informed
- ☐ Export received and verified; decision records retained per local retention rules
- ☐ Vendor deletion certificate on file: Data Steward _____ Date _____
- ☐ Access closed item by item; register entry closed; one page of lessons filed with the procurement templates

Artifact 6b: Monitoring plan template

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0. One page per system, completed before go-live; the G3 gate checks it. Baselines come from shadow mode. Every row carries a named person and two thresholds; a row with a title instead of a name, or a threshold left “to be set,” fails the gate. Quantitative rows carry numbers in both threshold columns. Qualitative rows (input data quality, source changes) carry the review action instead: “any occurrence: note and review at month end” in the investigate column, and the named condition that triggers the pause question in the act column.

System: _____ Tier: _____ Baselines from shadow period: _____ to _____					
Metric	Baseline	Cadence	Who looks (name)	Investigate at	Act at (pause question same day)
Volume per week, by site or channel		Weekly/Monthly			
Output distribution (rate per outcome class or score band)		Monthly			
Override and exception rate (Module 3.5; defaults: investigate below 1 percent or above 30 percent)		Monthly			
Unflagged-case audit findings (sample: 20 or 2 percent monthly)		Monthly			
Quarterly outcome reconciliation (silent failures found)		Quarterly			
Subgroup performance, or stratified audit where outcomes lag		Quarterly			
Appeals: count, upheld count, themes (three upheld in a quarter = incident question, automatic)		Monthly			
Input data quality (missing fields, malformed records, source changes)		Monthly			

Summary to System Owner: monthly, one page. Plan reviewed at every G4 and after every accepted major update. Prepared by _____ (Operational Lead) Date _____

Artifact 6c: Appeal notice template

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Licence: CC BY 4.0. Plain-language text for letters, portals, and counters, wherever a person meets a decision the system informed. Adapt the bracketed items; keep the reading level; translate into every language your institution serves people in. Check locally: where your law requires telling people an automated system was involved, or grants a right to human review, your notice must say so; this template is the floor. Conversational systems: replace the reference number with the session identifier, surface the challenge route inside the interface on every session, and publish this notice text where the chat is offered.

If you think this decision is wrong, you can ask us to look at it again.

A computer system helped us assess your case. A person is responsible for the decision, and a different person will review it if you ask.

How to ask: [channel: form, address, phone, counter], quoting reference [x]. You do not need to give a reason, though anything you tell us helps.

What happens next: we will confirm we received your request within [5] working days, and give you our answer, with reasons, within [20] working days.

Asking for a review will not be held against you and will not delay anything you are currently receiving.

[Delete or adapt where a decision has suspensive effects; state them plainly.]

References for this module

- Wong, A., et al. "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients." *JAMA Internal Medicine*, volume 181, issue 8 (2021), pages 1065 to 1070. (Full citation and the accompanying editorial in Module 4's references.)
- Ross, C. "Epic overhauls popular sepsis algorithm criticized for faulty alarms." *STAT News*, 3 October 2022. (Reports, from the developer's documents, the recommendation that the model be trained on a hospital's own data before clinical use, a changed definition of sepsis onset, and reduced reliance on antibiotic orders as a flag.)
- Wong, A., Currey, D., Schwinne, M., Park-Egan, B., Meyer, S., Gutting, A., Cao, J., Khan, S., Dantes, R., Pan, T., Buchman, T., Singh, K., Bhavani, S. V., Lyons, P. G., Sjoding, M. W., and Tarabichi, Y. "Multicenter Prospective Validation of an Updated Proprietary Sepsis Prediction Model." *JAMA Network Open*, volume 9, issue 2 (2026), article e260181. doi:10.1001/jamanetworkopen.2026.0181. (States that the revised model was released in 2022 and was fine-tuned on a site-specific historical training set at all study sites before implementation.)

Module 7: The implementation toolkit

This module puts every artifact in one place, each in its final, standalone form, and gives one person a thirty-day path from “we have no idea what AI we run” to a classified register, named owners, and the first controls live.

If you are starting from zero, walk the thirty-day path in 7.0; it names the artifact for each step. If you are mid-procurement, go straight to the artifact you need from the index below. Each artifact repeats its own instructions, carries its own licence line, and works without its parent module, though the parent module is where its reasoning lives. These are the authoritative versions: where wording here and in a module ever differ, this module governs. Every artifact is also published as a separate editable file (document or spreadsheet format); the front matter says where. Nothing here needs to be retyped. The artifacts define required information: reproducing the same information inside an existing governance or ticketing platform satisfies the framework. And every artifact, once adopted, carries a document control block: document version, prepared by, last updated, next review; the editable files include it, so add it when copying from this page.

The index.

#	Artifact	Use it when	Parent
7.1	One-page orientation	Briefing leadership; asking for the four role appointments	Module 1
7.2	Risk classification worksheet	Classifying any AI use case, new or already live	Module 2
7.3	Vendor due diligence checklist	Scoring vendors before signature	Module 4
7.4	Contract clause library	Negotiating the contract; briefing local counsel	Module 4
7.5	Governance RACI and role definition pack	Appointing the four roles; writing the appointment letters	Module 3
7.6	Pre-deployment readiness pack (gate G2 and G3 cover document, with the data checklist)	The readiness and go-live gates	Modules 5 and 6, new cover document
7.7	Stage-by-stage lifecycle checklist	Running a deployment start to finish	Module 6
7.8	Monitoring plan template	Before go-live, and at every G4 review	Module 6
7.9	Appeal notice template	Publishing the challenge route to affected people	Module 6
7.10	Incident response template	The moment the incident question is asked	Module 6, new form
7.11	System register	Always; it is the living index of everything above	New, defined here

Two artifacts are new in this module: the incident response template (7.10), which Module 6 defined in prose and this module turns into a form, and the system register (7.11). One is new as a cover document: the readiness pack (7.6) binds the Module 5 data checklist and the Module 6 validation evidence into the single document the Independent Reviewer signs.

7.0 The thirty-day path

For the institution that has systems already running, no governance around them, and one person (you) with part of their time. Each step names its artifact and its output. The path assumes roughly two working days per week (a framework default); where you have less, the weeks stretch and the order holds. It starts with a short written directive from leadership adopting the framework and instructing departments to answer week 1’s question (the sign-off block on Artifact 7.1 exists for this).

Week 1: find everything.

- Ask IT for the software asset inventory and the identity provider’s list of connected applications.

- Review accounts payable records for AI-related line items.
- Network and web logs can complete the picture where your policy and local law allow their use for this purpose (check locally).
- Ask every department head one question by email (“which tools in your team produce text, scores, predictions, or decisions automatically? Include free ones”) and send a reminder a week later.

Open the system register (7.11) and give every finding a row, even the embarrassing ones. Output: a register with every AI use in the institution on it, owner column empty.

Week 2: classify everything. Run the risk worksheet (7.2) on each row, quickest first, and get each classification confirmed by a second person per the worksheet's signature rule. Anything that hits the escalation list stops and routes to legal review. Output: every register row carries a tier and two signatures.

Week 3: get owners named. Send leadership the one-page orientation (7.1) with your register attached; the meeting you are asking for has one agenda item: appointing owners for every classified system per the Module 2 load table:

- All four roles for each Tier 2 system.
- A System Owner, Operational Lead, and Data Steward for Tier 3.
- A portfolio System Owner and an Operational Lead for the Tier 4 tools.
- The Independent Reviewer arrangement (the role pack in 7.5 makes the appointments a signature).

Write the acceptable-use note for each general-purpose tool (rule 1 on the worksheet). Output: no classified row with an empty owner column; users of general tools told what the tools are for.

Week 4: put the first controls on the riskiest thing. Take the highest-tier system that is live today. With its new Operational Lead: draft its monitoring plan (7.8) from whatever baseline data exists, publish the appeal notice (7.9) where its decisions land, and pull its contract to check three things against the clause library (7.4): can we exit, can the vendor update silently, and who owns our data. For a live Tier 2 system, gaps in exit, update control, or data rights do not wait: the System Owner decides within 30 days (a framework default) between opening renegotiation with the vendor now and signing a written, dated risk acceptance that leadership sees. Gaps on lower-tier systems go on a renewal fix list, dated for the next negotiation. Output, day 30: a one-page status to leadership: what we run, what tier, who owns it, what the first monitoring says, and the contract gaps with dates.

The path deliberately postpones full validation of already-live systems (schedule it as 6.2 late validation at each system's first G4 date) and the full procurement machinery, which begins the day the next purchase does. Scale notes: a five-person organisation runs the same path with the roles doubled up, keeping the one separation (the person who proposes a system never confirms its classification or signs its gates alone); a very large estate runs weeks 1 and 2 per division into one consolidated register, and the register's monthly review becomes the portfolio governance meeting.

7.0.1 What the completed toolkit evidences

Filled in, these artifacts are the paper trail an auditor, regulator, board, or ministry asks for. The mapping runs through the instruments named in Module 1:

Evidence the toolkit produces	What it evidences, at concept level
The register (7.11) and signed worksheets (7.2)	Risk classification: the EU AI Act's risk logic; NIST AI RMF Map
The RACI, role appointment letters (7.5), and gate signatures (7.6, 7.7)	Accountability and human oversight: the Act's deployer duties; AI RMF Govern; ISO/IEC 42001 leadership clauses
The due diligence checklist (7.3), clause outcomes (7.4), and data checklist (7.6a)	Supplier and data governance: the Act's documentation exchange; ISO/IEC 42001 supplier and data expectations
The monitoring plans (7.8), incident records (7.10), and appeal logs (7.11)	Operation and improvement: the Act's monitoring duties; AI RMF Measure and Manage; ISO/IEC 42001 operation, performance evaluation, and improvement

7.1 The one-page orientation (from Module 1)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. Purpose: a single page for a director, board member, minister, or executive whose institution is adopting this framework. It is designed to be sent ahead of the meeting where you ask for the four appointments at the bottom.

AI governance for institutions that deploy: one page

The situation. Our institution buys and runs AI systems; it does not build them. Almost all AI governance guidance is written for builders. This framework is an operating manual for our seat: what to demand from vendors, who is accountable internally, what to check before a system goes live, what to watch afterwards, and how to switch off.

The framework we are adopting. The Open AI Governance and Deployment Framework (Version 1.0), a public document published under an open licence. It maps onto the EU AI Act, the NIST AI Risk Management Framework, ISO/IEC 42001, the OECD AI Principles, and the UNESCO Recommendation on the Ethics of AI, so work done under it is evidence of movement towards recognised standards, though the framework certifies compliance with none of them. It is not legal advice, and it prescribes where legal and specialist review is required.

The five commitments it runs on.

1. Accountability names a person: every system has one named owner with the authority to stop it.
2. Evidence comes before trust: no vendor claim without a document; no high-stakes go-live without a test on our own data.
3. Oversight carries real authority: the humans reviewing AI outputs can override them, and have the training and time to do it.
4. Watching continues after go-live: defined metrics, a defined reviewer, defined thresholds for action, and a channel for affected people to challenge a decision.
5. We can always leave: every contract includes exit and data return; every system has a switch-off plan.

What this asks of leadership.

1. Appoint four roles (existing staff hold them alongside their jobs, with the monthly time cost stated per role): a System Owner per AI system, an Operational Lead, a Data Steward, and an Independent Reviewer for go-live sign-offs.
2. Require the risk classification worksheet for every AI use case, existing and proposed, including free tools.
3. Require the procurement gate: no AI contract is signed until the vendor due diligence checklist passes for the system's risk tier.
4. Receive the monitoring summary: one page per month on every high-stakes system, from the System Owner.

What it costs. Staff hours, stated per role in the framework, and paid specialist help at the points named in Module 1.4: legal review, data protection advice, security testing, adversarial testing, and domain review. What it prevents: buying systems that do not work on our population, contracts we cannot exit, silent degradation in service, and standing unprepared in front of a regulator, an auditor, or an affected person.

Directive and designations. By signing, leadership adopts this framework for the institution and mandates the thirty-day path (Module 7), including the instruction to departments to answer its first-week question.

Role	Name	Date
System Owner for [system]		
Operational Lead		
Data Steward		
Independent Reviewer arrangement		

Authorised by (name, title, signature): _____ Date: _____

Prepared by: _____ For: _____ Date: _____

7.2 Risk classification worksheet (from Module 2)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. One worksheet per use case, including free tools and systems already live. It is designed to be completed in under an hour once the system description (Module 4, section 4.2.1) is in hand. It is an operating tool and it is not legal advice; where a law classifies your system, the law's classification governs.

Use case being classified

System and vendor	
The use case, in one sentence (what decision or output, for whom)	
Users of the output	
People affected by the output	
Expected volume (outputs per week)	
Model or product version reviewed	
Deployment location(s) and sites	

Step 1: Escalation list. Tick any that apply.

- ☐ Scores or ranks people's general trustworthiness or social behaviour across contexts
- ☐ Biometric identification in public spaces, or biometric categorisation of people
- ☐ Infers emotions or protected characteristics about employees, students, or the public
- ☐ Adverse decisions about people, fully automated, with no individual human review and no appeal
- ☐ Inputs include protected characteristics or close proxies, in a system touching decisions about people (rule 5)
- ☐ Known to be prohibited or restricted by a law that applies to us (state which), or legal status unknown (if unknown, do not decide: route this worksheet to legal review before classifying further): _____

Any tick: **STOP. Tier 1. Route to legal counsel and a relevant specialist with this worksheet attached. Leadership sign-off required to proceed on any path.**

Step 2: Decision question. Does the system inform or make decisions that can be adverse to an identifiable person (care, money or credit, benefits, housing, education, employment, essential services, legal standing, liberty)?

- ☐ Yes → **Tier 2.** Go to Step 4. ☐ No → Step 3.

One sentence of reasoning: _____

Step 3: Dependence question. Do the outputs steer operations that people depend on, without deciding about any identifiable person?

- ☐ Yes → **Tier 3.** ☐ No → **Tier 4.** Go to Step 4.

One sentence of reasoning: _____

Step 4: Modifiers. These move the case up only.

- **Autonomy:** will individual outputs take effect without prior human review? ☐ No ☐ Yes → Tier 3 or 4 moves up one tier; at Tier 2 this is Step 1 territory: STOP and escalate.
If human review is claimed, show the arithmetic: (outputs per week _____ ÷ reviewers _____) ÷ working days per week _____ = _____ reviews per reviewer per day. Is that real review? ☐ Yes ☐ No → treat as no prior human review.
- **Detectability:** if an output were wrong, would anyone notice within days? State the mechanism that would notice it: _____
☐ Yes ☐ No → Tier 3 or 4 moves up one tier; at Tier 2, the strict Module 6 path is mandated (shadow mode before go-live; monitoring designed around unnoticed failures).

Result

Final tier	
Governance load acknowledged (attach or reference the tier table, Module 2.3)	☐
Acceptable-use note written (general-purpose tools only, rule 1)	☐ attached
Re-classification triggers acknowledged (rule 4), plus any system-specific trigger, stated: _____	☐
Next scheduled review date (12 months Tier 2 and 3; 24 months Tier 4)	

Signatures

	Name	Role	Date
Classified by			
Confirmed by (System Owner for Tier 2; a peer outside the proposing team otherwise)			

File this worksheet in the system register (Module 7, Artifact 7.11).

7.3 Vendor due diligence checklist (from Module 4)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. This checklist is an operating tool and it is not legal advice.

How to score. For each item, enter: **0** (absent, refused, or answered only verbally), **1** (partial, or promised in writing but not yet delivered), **2** (delivered in a written document that will be attached to the contract). Items marked ★ are disqualifying items: a score of 0 on any ★ item fails the checklist for a Tier 2 purchase regardless of the total.

Pass thresholds.

Tier (from the Module 2 worksheet)	Threshold
Tier 2 (high stakes)	No ★ item at 0, and at least 80 percent of all available points across sections A to F
Tier 3 (operational)	No 0 on items A1, B1, D1, D3, F1, and on E3 where Section E applies, and at least 60 percent of available points on sections A, B, D, F
Tier 4 (low stakes)	Items D2, F1, F3 all at 2. Other sections optional

Section A: System documentation

#	Item	Score
A1 ★	System description document: intended use, excluded uses, inputs, outputs, system type	
A2	General training data description: time period, geography, languages, population coverage	
A3 ★	Written limitations statement: degradation conditions, failure modes, untested settings	
A4	Instructions for use suitable for training your oversight staff (Module 3)	

Section B: Performance evidence

#	Item	Score
B1 ★	Every performance claim in writing, each with metric, dataset, and population described	
B2 ★	Written answer to the population question: evidence of measurement on a population like yours, or an honest statement that none exists	
B3	Subgroup performance breakouts relevant to your setting, or a written statement of what was not measured	
B4	Method described in enough detail to be checked	
B5 ★	Written agreement to acceptance testing on your data, against criteria to be fixed in the contract	

Section C: Vendor interview (section 4.4)

#	Item	Score
C1	Interview held, answers recorded verbatim, record in the procurement file	
C2	Written follow-ups received for every question deferred in the interview	
C3	Vendor confirmed (question 16) that its answers will attach to the contract	
C4	Reference calls held with at least two comparable customers, notes on file	

Section D: Data, security, and logging

#	Item	Score
D1 ★	Written statement of storage, processing, and backup locations, by country	
D2 ★	Written answer on training use of your data and your users' inputs, with an off switch	
D3 ★	Audit logging confirmed in writing: what is logged, access, retention, export format	
D4	Security certifications provided with scope statements covering this product	
D5	Named subcontractor list with access and locations	
D6	Incident history for this product (three years) and notification commitment in hours or days	

Section E: Generative tools only (skip for purely predictive systems)

#	Item	Score
E1	Measured factual-error rate with method	
E2	Prompt injection testing evidence	
E3 ★	Training exclusion for your users' inputs, available and confirmed in writing	
E4	Per-user logging, exportable by your institution	

Section F: Contract posture

#	Item	Score
F1 ★	Exit and portability clause agreed: formats, days, fixed cost, survival	
F2 ★	Audit rights clause agreed	
F3 ★	Update and change control clause agreed: notice days, changelog, test window, version hold	
F4	Documentation-as-warranty clause agreed; entire-agreement wording checked accordingly	
F5	Breach notification hours and incident notification days written into the contract	
F6	Liability cap written as a currency figure; compared against the estimated cost of the worst realistic failure of this system; data breach liability confirmed as sitting outside or above the general cap; local counsel review completed and dated	
F7	Named termination triggers in the contract	

Red flag check (section 4.5). Number of red flags present: _____. One flag: purchase paused until resolved in writing. Two or more on a Tier 2 purchase: process ends with this vendor.

Result and sign-off

System and vendor	
Risk tier (attach the Module 2 worksheet)	
Total score / available	
★ items at 0	
Result (pass / fail / paused)	
Completed by (name, role, date)	
Approved at procurement gate (name, date)	

7.4 Contract clause library (from Module 4)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0.

Read this before using any clause. These clauses are drafting guidance written in plain language. They are not legal advice, and contract law differs by country. A clause that works in one jurisdiction may need different words, or may be unenforceable, in another. Before signature, have a lawyer qualified in your jurisdiction adapt and review the final text. Bring them this library and the vendor's draft together; the comparison will focus their work and reduce their fee. In these clauses, "the Institution" is your organisation, "the Vendor" is the seller, and "the System" is the AI system being purchased. Bracketed items in [square brackets] are numbers or lists you must fill in; a clause with an empty bracket is not finished. Each clause carries a marker: **(M)** mandatory, where walking away beats signing without them for a Tier 2 system; **(R)** recommended, concede only knowingly and in writing.

Clause 1: Documentation as warranty (M) The documents listed in Schedule [x] (including the system description, the limitations statement, the performance evidence, and the Vendor's written tender and interview answers) form part of this contract. The Vendor warrants that they are accurate and complete in all material respects. A material inaccuracy in any of them is a breach of this contract. *What this protects you from:* the sales claims evaporating at the moment you need them; an "entire agreement" clause quietly deleting everything you were told before signing.

Clause 2: Acceptance testing (M) The System will not enter live use, and the payment milestone in Schedule [x] will not fall due, until the Institution has tested the System on its own data against the acceptance criteria in Schedule [x] and the System has met them. The Vendor will provide the cooperation, configuration, and technical support needed to run the test. If the System fails, the Institution may require remediation and a retest, renegotiate, or terminate under Clause 11 with a refund of amounts paid for the System. The test will begin within [x] days of signature and complete within [x] days; if the Vendor's cooperation is not provided in time to meet those dates, the milestone does not fall due and the Institution may terminate under Clause 11. Acceptance or rejection is given in writing by the Institution's System Owner. Where fees are paid by subscription rather than milestones, the first [x] months' fees are refundable in full if the System fails acceptance. *What this protects you from:* discovering, as the buyers in section 4.1 did, that claimed performance and delivered performance are different systems.

Clause 3: Ongoing performance warranty (R) The Vendor warrants that, in live use, the System will perform within the ranges in Schedule [x], which states each metric by name (for example: accuracy, error rate by subgroup, alert or exception rate, response time), each with its measurement method. If measured performance falls outside the warranted range and is not restored within [x] days of notice, the Institution may terminate under Clause 11. *What this protects you from:* the system that passes on day one and degrades quietly in service, with no lever attached to the decline.

Clause 4: Audit rights (M) The Institution, or an independent auditor it appoints under confidentiality obligations, may audit the System's logs, the documentation, and the Vendor's compliance with this contract's data and security commitments: at least once per year, and additionally after any incident or suspected breach. The Vendor will provide access and cooperation, with audit logs and compliance documentation made available within [72] hours of a written request. The Institution gives [x] days' notice for routine audits and none for incident audits; the auditor is bound in writing to confidentiality; audits are conducted so as not to disrupt the service; and each party bears its own costs, except that audits beyond the first in any year are at the Institution's cost. *What this protects you from:* commitments that are true only in the sales meeting.

Clause 5: Update and change control (M) The Vendor will give the Institution at least [30] days' written notice of any change to the model or to System behaviour, with a written description of what is changing and why. The Institution may test the change before it is applied to its live environment and may decline it, remaining on its current version, for up to [x] months. Emergency security fixes may be applied immediately, with notice and description to follow within [x] days. If an applied change degrades the System against Schedule [x], the Vendor

will restore the Institution's prior version (rollback) within [x] days of notice, at no charge. *What this protects you from:* silent updates: the validated system being replaced, unseen, by a system nobody validated.

Clause 6: Data ownership and permitted use (M) All data the Institution or its users provide to the System, and all outputs generated from it, remain the property of the Institution. The Vendor will use this data only to provide the contracted service. The Vendor will not use it to train, improve, or develop any model or product, and will not sell, share, or derive products from it, in original, aggregated, or anonymised form, without a separate written agreement signed by the Institution. *What this protects you from:* your institution's records, and your users' inputs, becoming someone else's asset by default. *Check locally:* who can own AI-generated output, and what rights exist in it, varies by jurisdiction; have counsel adapt the ownership words to your law. The clause's enforceable core is the restriction on the Vendor's use, which does not depend on ownership doctrine.

Clause 7: Data location and subcontractors (M for Tier 2) The Institution's data will be stored, processed, and backed up only in the countries listed in Schedule [x], and accessed only by the Vendor personnel and the subcontractors listed in Schedule [x]. Changes to either list require [x] days' advance written notice and, for the countries list and (for Tier 2 systems) the subcontractor list, the Institution's written consent. The Vendor will impose the obligations of this contract on every subcontractor and remains fully responsible for their performance. *What this protects you from:* your data migrating to jurisdictions, and into hands, you never evaluated; promises that stop at the vendor's front door.

Clause 8: Security and breach notification (M) The Vendor will maintain the security controls in Schedule [x] throughout the contract. The Vendor will notify the Institution of any security breach affecting the Institution's data within [x] hours of discovery, with a written account of scope, cause, and remediation to follow within [x] days. *What this protects you from:* learning about a breach after your own legal notification deadlines have already expired. Set the hours after checking your local breach notification law.

Clause 9: Product incident notification (R) If the Vendor becomes aware of a material defect, error, or safety problem in the System or in the same product deployed elsewhere, it will notify the Institution within [x] days, with its assessment of relevance to the Institution's deployment. *What this protects you from:* being the last customer to learn what the vendor already knows.

Clause 10: Audit logging (M for Tier 2) The System will log, at minimum: each input received, each output produced, the date and time, the model version, and the identity of the requesting user or process. Logs will be retained for [x] months, remain accessible to the Institution throughout, and be exportable by the Institution in the documented format in Schedule [x] at no additional charge. Logs will be protected against alteration, and any gap or edit in the log record is itself notifiable under Clause 8. *What this protects you from:* being unable to answer an affected person, an auditor, or your own incident investigation, because the record of your system's decisions belongs to someone else.

Clause 11: Termination triggers (R) In addition to any other rights, the Institution may terminate this contract if: the System fails acceptance testing under Clause 2; the System fails the performance warranty under Clause 3; the Vendor applies a change in breach of Clause 5; the Vendor refuses or obstructs an audit under Clause 4; the Vendor materially breaches Clauses 6, 7, or 8; or the Vendor announces withdrawal or end of support of the System or of the model version relied on. Termination under this clause entitles the Institution to a refund of fees paid for undelivered or unusable periods. *What this protects you from:* a contract where every exit requires proving "material breach" from a standing start, in a dispute, at your expense.

Clause 12: Exit and portability (M) On termination or expiry for any reason, the Vendor will: (a) export and deliver to the Institution all Institution data, configurations, and audit logs, in the open, documented formats in Schedule [x], within [x] days, for the fixed fee in Schedule [x]; (b) provide transition assistance for [x] months at the rates in Schedule [x]; and (c) then permanently delete all Institution data from its systems and its subcontractors' systems and certify the deletion in writing. The obligations in this clause survive termination and apply even while any dispute between the parties is unresolved. The export format will be validated during acceptance testing. Where the System is critical to continuing operations, the parties will record in Schedule [x] an escrow or continuity arrangement covering Vendor insolvency or withdrawal of the product. *What this protects you from:* lock-in, which is the quiet tax on every renewal; and a vendor holding your data hostage during a dispute.

Clause 13: Intellectual property indemnity (M for generative systems) The Vendor will defend and indemnify the Institution against third-party claims that the System, or output generated by the System when used as documented, infringes intellectual property rights, and will at its own cost secure the Institution's right to continue use, modify the System to be non-infringing, or refund and assist exit under Clause 12. *What this protects you from:* your institution answering for what the vendor's training data contained.

Clause 14: Price protection and product continuity (R) The Vendor's charges for renewal terms will not increase by more than [x] percent per year, or by more than [named index], whichever is lower. The Vendor will give at least [x] months' written notice before withdrawing, discontinuing, or ending support of the System or of the model version relied on, and throughout that notice period exit assistance under Clause 12 remains available at the rates in Schedule [x]. *What this protects you from:* a dependency being repriced once switching has become expensive, and a product disappearing faster than you can leave it.

7.5a Governance RACI (from Module 3)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. R = Responsible (does the work). A = Accountable (answers for it). C = Consulted. I = Informed. Adjust to your structure and file the adjusted version in the system register. Tier note: this chart shows the Tier 2 allocation. At Tier 3, the Independent Reviewer column falls away (Module 3.2). At Tier 4, the Operational Lead holds the A for the G1 and G3 gate rows, per the gate table in Module 3.3.

Activity	System Owner	Operational Lead	Data Steward	Independent Reviewer
Risk classification worksheet (Module 2)	A (confirms Tier 2)	R	C	C (confirms below Tier 2, if peer)
Tender demands and vendor interview (Module 4)	A	R	C (data demands)	I
Due diligence checklist and G1 signature	A	R	C	I
Contract signature readiness (clauses, counsel review)	A	C	C	I
Pre-deployment data checklist (Module 5)	I	C	A and R	I
Validation on own data (Module 6)	A	R	C	I
G2 readiness signature	I	C	R (data part)	A (whole)
Oversight design (3.4 one-pager)	A	R	C	C (checks at G3)
G3 go-live signature	A	R (evidence pack)	C	A (joint)
Monthly monitoring review and summary (Module 6)	I (receives)	A and R	C (data metrics)	I
Override, reason, and blind-sample review (3.5)	I	A and R	I	I
Pause decisions (3.6)	A (beyond 48 hours)	R (first 48 hours)	I	I
Incident declaration and response (Module 6)	A	R	C (data incidents: R)	I
Vendor update acceptance and re-test (Module 6)	A	R	C	I
Grievance and appeal responses (Module 6)	A	R	C	I
G4 continued-operation review	A	R	C	I
Decommissioning and data return (Modules 4 and 6)	A	R	R (deletion certificate)	I

7.5b Role definition pack (from Module 3)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. Four half-page definitions, written to be pasted into an appointment letter or job description. Replace bracketed items. Time budgets are the framework defaults from Module 3.2; adjust and record. Tier note: at Tier 4, the G1 and G3 gate signatures sit with the Operational Lead (Module 3.3); adjust the two definitions accordingly when appointing for a Tier 4 portfolio.

System Owner, for [system name]

Purpose. You are accountable for this system's behaviour in service. You are the person who answers for it, to leadership, to auditors, and to affected people, and you hold the authority that accountability requires. **Authority.** You may halt the system at any time on your own decision. You approve or refuse the procurement gate (G1), the go-live gate (G3, jointly with the Independent Reviewer), and the periodic continued-operation gate (G4). You own incident decisions beyond the Operational Lead's 48-hour pause authority. **You sign.** The due diligence checklist (G1), the go-live checklist (G3), continued-operation reviews (G4), and the register entry carrying your name. **You may refuse.** Any gate where a claimed item lacks a document. Refusals are recorded, and recorded refusal is a protected act. **You receive.** The monthly monitoring summary, every pause and bypass record, and every incident declaration. **Time.** About 4 hours per month in steady operation for a Tier 2 system; 10 to 15 hours one-time across procurement and go-live. **Seniority test.** If you would need permission to stop this system, the role sits too low. Hand it up.

Operational Lead, for [system name]

Purpose. You run this system day to day and you are the first pair of eyes on everything it does in service. **Authority.** You may pause the system for up to 48 hours on your own decision. You decide pause requests within 4 working hours, with one recorded line of reasoning. You accept or defer vendor updates into the test window (Module 6) and keep the acceptable-use note current. **You do.** The monitoring review and the monthly summary; the override-rate, written-reason, and blind-sample countermeasures (Module 3.5); reviewer training records; the oversight design one-pager (Module 3.4) for the Independent Reviewer's G3 check; escalation intake. **You may refuse.** To lift a pause you imposed until the System Owner has reviewed it. **Time.** About 8 hours per month in steady operation for a Tier 2 system; 20 to 30 hours one-time through validation and rollout.

Data Steward, for [system name]

Purpose. You control what data this system receives, produces, and sends onward, and nothing about that changes without you. **Authority.** You approve or refuse the pre-deployment data checklist (Module 5), and any subsequent change to data sources, fields, destinations, or storage locations. Your refusal blocks G2. **You sign.** The data section of G2; deletion certificates at decommissioning. **You are the contact for.** Data questions from auditors, regulators, affected people exercising data rights (check locally which rights apply), and the vendor. **Time.** About 2 hours per month in steady operation; 8 to 12 hours one-time, mostly the Module 5 checklist.

Independent Reviewer, for [system name]

Purpose. You are the outside check with the right to refuse the signature, at the two gates where enthusiasm and momentum are least reliable: readiness (G2) and go-live (G3). **Independence.** You sit outside the team deploying this system and had no part in proposing or procuring it. If your department and theirs review for each other, the arrangement is recorded in the register. **Your method, per gate.** Take the gate checklist. For every item marked done, ask for the document. Read the validation results against the criteria that were agreed before testing, and check the oversight design one-pager against the four elements in Module 3.4. Where a document is missing or a criterion unmet, refuse the signature and record why. You are not asked to judge the technology; you are asked to confirm the evidence exists. **You may refuse.** The signature. That is the role. **Time.** About 8 hours per gate, twice per system.

7.6 Pre-deployment readiness pack (gates G2 and G3)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. The single cover document for the readiness gate (G2) and the go-live gate (G3), designed for the Independent Reviewer's method (Module 3): every item claims a document, and the reviewer asks for each document. The pack is this cover sheet plus its attachments; an item without its attachment fails the gate. The full data checklist (7.6a, below) is attachment A.

System: _____ **Tier:** _____ **Register row:** _____

Gate G2: Readiness

#	Item claimed	Attachment	Present
A	Pre-deployment data checklist complete, local-advice lines signed where required	7.6a, completed	☐
B	Acceptance criteria as fixed in the contract schedule (before testing)	Contract schedule extract	☐
C	Validation run and results against those criteria, subgroups included (Module 6.2)	Results record	☐
D	Comparison against current practice recorded	In results record	☐
E	Frozen test set stored under Data Steward control, excluded from vendor training in writing	Storage note and contract clause reference	☐

G2 signatures: Data Steward (item A) _____ Independent Reviewer (whole) _____ Date _____

Gate G3: Go-live

#	Item claimed	Attachment	Present
F	Shadow mode completed to its minimums; baselines captured	Shadow report	☐
G	Pilot completed against written scope and stop criteria, decision recorded	Pilot paragraph and decision	☐
H	Oversight design one-pager: authority, capacity (minutes per case from shadow), information, consequence (Module 3.4)	One-pager	☐
I	Reviewers trained on the limitations statement and validation failures; records kept	Training record	☐
J	Monitoring plan complete: every row a name and two thresholds	7.8, completed	☐
K	Appeal channel live; notice published where decisions land	7.9 as published	☐
L	Incident template staged; severity definitions aligned with the contract	7.10, blank but configured	☐
M	Register row updated: version, dates, roles	7.11 extract	☐

G3 signatures: System Owner _____ Independent Reviewer _____ Date _____

7.6a Pre-deployment data checklist (from Module 5, attachment A to the readiness pack)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. Completed and signed by the Data Steward before the readiness gate (G2, Module 3); an incomplete checklist blocks the gate. It is an operating tool and it is not legal advice; items marked “local advice” require answers from a person qualified in your jurisdiction, and their name and date go on the line.

System

System and vendor	
Risk tier (attach the Module 2 worksheet)	
Vendor data specification received (Module 4.2.1)	☐ attached

Section A: Inventory. One row per field or field group leaving the institution. Attach continuation sheets as needed.

Field	Sensitivity class (5.2)	Needed for function? (vendor’s written answer)	Withheld instead?	Regulator, if regulated data

Section A2: Unstructured sources (generative deployments; one row per source)

Source	Worst sensitivity class realistically inside	Scrubbing or redaction control, verified on a sample by	Needed for function?	Derived stores (embeddings, caches, memory) covered by location, retention, deletion answers?

Section B: Lawful basis (local advice for Tier 2 and any sensitive personal data)

- ☐ Legal basis for this use identified and recorded: _____
- ☐ Purpose compared with collection purpose; notice or consent steps identified: _____
- ☐ People in the data informed, where required; how: _____
- ☐ Formal risk or impact assessment required? ☐ No ☐ Yes → completed and attached
 - Local advice by (name, qualification, date): _____

Section C: Location and access

- ☐ Storage countries, in the contract schedule: _____
- ☐ Processing countries: _____ Backup countries: _____
- ☐ Countries from which vendor or subcontractor staff can access: _____
- ☐ Transfer restrictions checked locally; instrument named: _____
- ☐ Region setting configured and recorded, where the product offers one
- ☐ Governing law of vendor and compelled-disclosure question put in writing to vendor and adviser; answers filed

Section D: AI-specific security (5.4)

- ☐ Named-account access only; access list at go-live filed; quarterly review scheduled

- ☐ Injection testing evidence received (systems reading untrusted text); independent security test commissioned for public-facing systems (name, date): _____
- ☐ Training use of our data: excluded in writing, or separately agreed with leadership signature attached
- ☐ Session and customer separation described by vendor in writing; canary test planned for validation
- ☐ Vendor access bounded in contract; access log review scheduled into annual audit
- ☐ Log export confirmed; monthly query-volume review by the Operational Lead
- ☐ Decision systems: influential-inputs statement received and gaming metric added to the monitoring plan

Section E: Minimisation and retention

- ☐ Field set reduced to the vendor specification's justified minimum; withholdings recorded in Section A
- ☐ Any "anonymised" or "aggregated" claim: method in writing, taken to local advice where sensitive
- ☐ Retention, as dates, per flow, in the contract schedule: _____
- ☐ Deletion or return at end, with written certification, per Module 4 Clause 12

Section F: Special categories (5.6), where applicable

- ☐ Health data: regulator guidance identified; approvals or ethics review obtained where required
- ☐ Children's data: lawful basis answered by local advice; vendor experience recorded in due diligence
- ☐ Financial data: supervisor's outsourcing and cloud rules mapped; notifications or approvals done
- ☐ Government administrative data: statutory confidentiality checked; in-country or on-premises default applied or departure justified in writing

Sign-off

	Name	Role	Date
Completed and signed		Data Steward	
Local advice items confirmed (where applicable)			

File with the G2 gate record in the system register (Module 7).

7.7 Stage-by-stage lifecycle checklist (from Module 6)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. One per system; items initialled as completed; gate rows signed per Module 3. Durations are Tier 2 defaults.

Stage 1: Validation (2 to 6 weeks)

- ☐ Acceptance criteria fixed in the contract schedule before testing (accuracy, worst subgroup gap, operational limits)
- ☐ Test sample assembled: at least 300 cases, representative, subgroups covered, hard cases included
- ☐ Sample frozen under Data Steward control; excluded in writing from vendor training
- ☐ Run completed in a test configuration; results scored against pre-fixed criteria only
- ☐ Compared against current practice on the same cases
- ☐ Decision recorded: pass / remediate and retest / renegotiate scope / exit
- ☐ **GATE G2:** Data checklist (Module 5) and validation results signed: Data Steward _____ Independent Reviewer _____ Date _____

Stage 2: Shadow mode (minimum 4 weeks and 200 outputs; 6 to 8 weeks if low detectability)

- ☐ Outputs recorded, nobody acting on them; comparison method defined first
- ☐ Live alert or exception rate measured against the operational criterion
- ☐ Real minutes-per-case review cost measured; reviewer time budget set (Module 3.4)
- ☐ Monitoring baselines captured for every Artifact 6b metric

Stage 3: Pilot (4 to 12 weeks)

- ☐ One-paragraph pilot definition written first: scope, duration, success criteria, stop criteria with numbers
- ☐ Oversight design one-pager in force (Module 3.4); reviewers trained on limitations and validation failures
- ☐ Appeal channel live and notice published (Artifact 6c)
- ☐ Incident plan and monitoring plan (Artifact 6b) in force
- ☐ Pilot decision recorded either way
- ☐ **GATE G3 (full go-live):** System Owner _____ Independent Reviewer _____ Date _____

Stage 4: Operation (standing)

- ☐ Monthly monitoring review and summary to System Owner; weekly glance if low detectability
- ☐ Monthly unflagged-case audit (20 cases or 2 percent); quarterly outcome reconciliation
- ☐ Appeals log read monthly beside the metrics
- ☐ Every update notice through the 6.7 routine; change log current; frozen test set or prompt set re-run on model changes
- ☐ **GATE G4 (every 12 months and on trigger):** revalidation passed, roles filled, thresholds reviewed: System Owner _____ Date _____

Stage 5: Decommissioning (2 to 8 weeks)

- ☐ Trigger and decision recorded
- ☐ Fallback confirmed working and staffed before switch-off
- ☐ Users, operators, and (where relevant) affected persons informed
- ☐ Export received and verified; decision records retained per local retention rules
- ☐ Vendor deletion certificate on file: Data Steward _____ Date _____
- ☐ Access closed item by item; register entry closed; one page of lessons filed with the procurement templates

7.8 Monitoring plan template (from Module 6)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. One page per system, completed before go-live; the G3 gate checks it. Baselines come from shadow mode. Every row carries a named person and two thresholds; a row with a title instead of a name, or a threshold left “to be set,” fails the gate. Quantitative rows carry numbers in both threshold columns. Qualitative rows (input data quality, source changes) carry the review action instead: “any occurrence: note and review at month end” in the investigate column, and the named condition that triggers the pause question in the act column.

System: _____ Tier: _____ Baselines from shadow period: _____ to _____					
Metric	Baseline	Cadence	Who looks (name)	Investigate at	Act at (pause question same day)
Volume per week, by site or channel		Weekly/Monthly			
Output distribution (rate per outcome class or score band)		Monthly			
Override and exception rate (Module 3.5; defaults: investigate below 1 percent or above 30 percent)		Monthly			
Unflagged-case audit findings (sample: 20 or 2 percent monthly)		Monthly			
Quarterly outcome reconciliation (silent failures found)		Quarterly			
Subgroup performance, or stratified audit where outcomes lag		Quarterly			
Appeals: count, upheld count, themes (three upheld in a quarter = incident question, automatic)		Monthly			
Input data quality (missing fields, malformed records, source changes)		Monthly			

Summary to System Owner: monthly, one page. Plan reviewed at every G4 and after every accepted major update. Prepared by _____ (Operational Lead) Date _____

7.9 Appeal notice template (from Module 6)

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. Plain-language text for letters, portals, and counters, wherever a person meets a decision the system informed. Adapt the bracketed items; keep the reading level; translate into every language your institution serves people in. Check locally: where your law requires telling people an automated system was involved, or grants a right to human review, your notice must say so; this template is the floor. Conversational systems: replace the reference number with the session identifier, surface the challenge route inside the interface on every session, and publish this notice text where the chat is offered.

If you think this decision is wrong, you can ask us to look at it again.

A computer system helped us assess your case. A person is responsible for the decision, and a different person will review it if you ask.

How to ask: [channel: form, address, phone, counter], quoting reference [x]. You do not need to give a reason, though anything you tell us helps.

What happens next: we will confirm we received your request within [5] working days, and give you our answer, with reasons, within [20] working days.

Asking for a review will not be held against you and will not delay anything you are currently receiving. [Delete or adapt where a decision has suspensive effects; state them plainly.]

7.10 Incident response template

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. Open one per incident, the moment the incident question is asked; a record opened and closed as “no incident” costs a page and is filed with the monitoring notes. Clocks and definitions are Module 6.6’s. It is an operating tool and it is not legal advice; external notification duties are governed by your local law, named on your Module 5 checklist.

Incident record

Incident id and system	
Opened by (name), date, time	
Type (Module 6.6 closed list): ☐ 1 wrong or harmful output reached a person (including via oversight bypass) ☐ 2 systematic error discovered ☐ 3 security event ☐ 4 vendor notification ☐ 5 unnotified model or behaviour change	
Severity: ☐ 1 wrong or harmful outputs are reaching people now ☐ 2 unavailable or unreliable but contained ☐ 3 degraded, suspected, or near-miss	

Contain (immediately / same day / next review, by severity)

Action taken (pause, channel stopped, single output held)	By	Date, time

Declare and inform: Operational Lead declared _____ System Owner informed _____ (within 4 working hours / 24 hours / 5 working days)

Assess (first written assessment within 5 working days): What happened, since when, evidenced by which logs: _____ Who was affected, how many: _____ Root cause, if known: _____

Notify

Who	Required by (law or contract clause; see Module 5 checklist)	Deadline	Done, date
Internal (leadership per severity)			
Affected persons			
Regulator(s)			
Vendor (or from vendor, Clause 8 or 9)			

Remediate (plan within 10 working days): Fix or restriction: _____ Decisions about people corrected, using the appeal machinery proactively (Module 6.6): _____

Restart (paused Tier 2 systems): frozen test set or relevant validation slice re-run and passed ☐ attached. System Owner signature _____ Date _____

Close and learn: cause, effect, and changes in three sentences: _____ New failure mode fed to reviewer training and the monitoring plan ☐ Register updated ☐ Closed by _____ Date _____

7.11 The system register

Standalone artifact. Open AI Governance and Deployment Framework, Version 1.0. Toolkit edition. Licence: CC BY 4.0. One living document for the whole institution: one row per AI use case, including free tools and pilots. It is the index to every other artifact (each row's documents live in that system's file), the first thing the thirty-day path creates, and the first thing to show an auditor. The Operational Lead of each system keeps its row current; a named person (default: the System Owner with the most Tier 2 systems; leadership names one where that ties) reviews the whole register monthly and reports rows with empty cells to leadership. Keep the register as a workbook: this table as the index tab, with gate records, incidents, and appeals on their own tabs linked by system name, because one row cannot hold a history.

Column	Entry
System and vendor	
Use case, one sentence	
Tier, worksheet date, next review date (7.2)	
System Owner / Operational Lead / Data Steward (names)	
Independent Reviewer arrangement, and reviewer name per gate record	
Gates passed, with dates (G1 to G4)	
Model or product version currently in use	
Last validation or revalidation, result, date	
Monitoring plan in force (7.8), last monthly summary date	
Appeal channel live (7.9), where published	
Contract: exit clause ☐ update control ☐ data rights ☐ (renewal fix list if gaps, with date)	
Status: ☐ proposed ☐ procurement ☐ validation ☐ shadow ☐ pilot ☐ live ☐ paused ☐ decommissioned	
Open incidents or fix-list items	
Framework version adopted (see Feedback and versioning)	

References for this module

This module introduces no new factual claims, statistics, or cases; every artifact's supporting evidence is cited in its parent module, and the standards mapping in 7.0.1 reuses only references verified in Modules 1 to 6.

Annex A: Full mapping to external standards

The concept-level mappings in each module, consolidated, with the specific references verified during drafting. The EU AI Act's obligations apply in phases, and readers should confirm current status; and this table shows direction of travel rather than certifying compliance with anything.

Module	EU AI Act (Regulation (EU) 2024/1689)	NIST AI RMF 1.0 (2023)	ISO/IEC 42001:2023	OECD AI Principles (2019, updated 2024) and UNESCO Recommendation (2021)
1: Foundations	Risk-based architecture of the Act as a whole; deployer obligations in Article 26	The four functions: Govern, Map, Measure, Manage	The management system as a whole (clauses 4 to 10: context, leadership, planning, support, operation, performance evaluation, improvement)	The values foundation the framework operationalises
2: Risk tiering	Risk categories; prohibited practices in Article 5; transparency duties in Article 50, applying from 2 August 2026 (telling people when they deal with an AI system, an emotion recognition system, or biometric categorisation; marking generated content; disclosing deepfakes, and AI-generated text published to inform the public on matters of public interest unless it has had human review and someone holds editorial responsibility)	Map function: context and risk categorisation	Risk assessment expectations within the management system	Proportionality: governance effort scaled to the risk of harm to people (the tier system)
3: Roles and oversight	Article 14 (design for effective human oversight: understand capacities and limitations, interpret output, decide not to use, intervene or interrupt); Article 26 (deployers assign oversight to persons with the necessary competence, training, and authority)	Govern function: accountability, roles, culture	Leadership and roles clauses	Accountability; UNESCO's human oversight and determination
4: Procurement	Article 11 (technical documentation) and Article 13 (instructions for use) as the provider-side documents a buyer demands; Article 26 deployer duties made performable by contract	Govern and Map: third-party risk	Supplier and third-party controls	Transparency and accountability, made contractual
5: Data, security, sovereignty	Article 10 (data and data governance for high-risk systems); national data protection law alongside and prevailing	Measure and Manage applied to data and security risk	Data management and impact assessment expectations; ISO/IEC 27001:2022 as the assumed general security baseline	Privacy; safety and security
6: Lifecycle	Article 26(5): deployers monitor operation on the basis of the instructions for use, and inform the provider and relevant authorities of serious incidents	Measure and Manage: test, monitor, respond, improve	Operation, performance evaluation, improvement, and change management clauses	Safety and accountability through the lifecycle; redress themes

Module	EU AI Act (Regulation (EU) 2024/1689)	NIST AI RMF 1.0 (2023)	ISO/IEC 42001:2023	OECD AI Principles (2019, updated 2024) and UNESCO Recommendation (2021)
7: Toolkit	The completed artifacts as the working evidence of the duties above	The four functions as filled-in documents	The records the management system requires	Documented evidence that the values were applied to each system, before and after go-live

Related instruments referenced in the text (check locally whether they apply to you): the GDPR (Regulation (EU) 2016/679): Article 22 (right not to be subject to solely automated decisions with legal or similarly significant effects, with exceptions), Article 33 (breach notification to the supervisory authority within 72 hours), Recital 26 (pseudonymised data is personal data). These are cited as examples of a family of national laws; your own jurisdiction’s instruments govern.

Annex B: Glossary

Plain-language definitions of technical terms used in this framework, alphabetically.

- **Acceptable-use note:** a one-paragraph statement of what a general-purpose tool may and may not be used for, shown to every user at account creation and first login.
- **Acceptance testing:** testing a system on your institution's own data, against criteria fixed in the contract before testing, with payment and go-live depending on the result.
- **Agentic system:** an AI system authorised to take actions on the deployer's behalf (executing transactions, sending messages, changing records) rather than only producing outputs for a human to act on. Handled in this framework through the autonomy rules in Module 2.
- **AI system:** software that produces scores, rankings, classifications, forecasts, or content by methods learned from data rather than rules written by hand. This framework covers predictive and generative kinds.
- **Anonymisation:** processing intended to make re-identification of people in data impossible. Hard for rich individual-level records; treat vendor claims of it as claims requiring the method in writing.
- **AUC (area under the curve):** a standard accuracy score for prediction systems, where 1.0 is perfect and 0.5 is a coin flip.
- **Audit log:** the system's own record of what happened: inputs, outputs, time, model version, and requesting user.
- **Audit rights:** the contractual right to inspect the vendor's compliance, logs, and documentation, directly or through an appointed third party.
- **Automation bias:** the human tendency to treat an automated cue as a substitute for looking at the evidence oneself; produces omission errors and commission errors.
- **Canary record:** a fabricated, harmless record planted in data to test whether information leaks where it should not.
- **Changelog:** a written list of what changed.
- **Commission error:** following a system's output even when available information contradicts it.
- **Data Steward:** the named role controlling what data a system receives, produces, and sends onward.
- **Decommissioning:** the ordered shutdown of a system: fallback first, people informed, data returned, deletion certified, records retained.
- **Deployer:** an organisation that puts an AI system into use, as distinct from the developer that built it.
- **Drift:** the slow movement of a system's output patterns away from their baselines without anyone changing anything, because the world stops matching the data the system learned from.
- **Due diligence:** the structured checking done before committing to a purchase.
- **Embedding:** a numerical representation of text or records that retrieval systems search over. Made from your data and treated as your data for location, retention, deletion, and training questions (Module 5).
- **Entire agreement clause:** a common contract clause stating that only the written contract counts; dangerous when it deletes the vendor's pre-contract claims, which is why those claims must be attached to the contract.
- **Escalation list:** the short list of uses (Module 2, Step 1) that no one classifies and deploys on their own authority.
- **Exit and portability:** the contractual guarantee that on leaving, the institution gets its data back in open formats, on a clock, at a fixed cost, with vendor deletion certified.
- **Framework default:** an operational number in this framework (an hour count, duration, sample size, threshold, or clock) chosen for workability rather than drawn from law or evidence; a recommended starting value to adopt, adjust, and record.

- **Frozen prompt set:** 30 to 50 representative tasks with notes on acceptable answers, kept unchanged and re-run on every generative model change.
- **Frozen test set:** the validation sample kept unchanged, under Data Steward control and excluded from vendor training, and re-run at every update and revalidation.
- **Gate:** a point where a named person signs a named document and the system does not move forward without the signature. This framework uses four (G1 to G4).
- **Generative system:** an AI system that produces content: text, images, or code. Includes chat assistants and drafting tools.
- **Hallucination:** a confident false statement produced by a generative system.
- **Incident:** any event on the closed list in Module 6.6, beginning with a wrong or harmful output reaching a person.
- **Independent Reviewer:** the named role that checks gate evidence and can refuse the signature, held outside the deploying team.
- **Liability cap:** a contractual ceiling on what a vendor can ever owe you, often a multiple of fees paid.
- **Minimisation:** sending the minimum field set the vendor's own data specification justifies, and recording what was withheld and why.
- **Model:** the learned component inside an AI system that turns inputs into outputs.
- **Omission error:** failing to act when a system stays silent even though other information says act.
- **Operational Lead:** the named role running a system day to day: monitoring, escalations, updates, and the countermeasures.
- **Override:** a reviewer's decision to depart from the system's output in an individual case.
- **Oversight:** a named person, with authority they can use, information they can act on, time to look, and accountability for what they decide.
- **Personal data:** anything relating to an identifiable person, including identifiers like account numbers, addresses, and device identifiers.
- **Pilot:** limited live use under a written scope, duration, success criteria, and stop criteria.
- **Predictive system:** an AI system that scores, ranks, classifies, or forecasts.
- **Procurement:** the whole process of buying: defining what you need, inviting offers, evaluating them, negotiating, and signing.
- **Prompt injection:** an attack where instructions hidden in text a system reads hijack its behaviour.
- **Pseudonymisation:** replacing direct identifiers with codes. The person remains identifiable to whoever holds the key, and most laws still treat the data as personal.
- **RACI:** a one-page chart assigning each activity a Responsible, Accountable, Consulted, and Informed role.
- **Register (system register):** the living document listing every AI use case, its tier, its owners, its gates, and its status.
- **Residency:** the countries where data is physically stored, processed, and backed up.
- **Revalidation:** re-running the validation exercise on refreshed data.
- **Risk tier:** the classification (1 to 4) that sets how much governance a use case must carry.
- **Sensitive personal data:** the categories of personal data given special protection by law, commonly health, ethnicity, religion, biometrics, sexual life, and criminal records, and often children's data. Check locally for your law's list.
- **Severity:** the three-level scale for incidents and support.
- **Shadow mode:** the system runs on live cases and its outputs are recorded, while humans decide exactly as before; outputs are compared against decisions and outcomes.
- **Silent update:** a change to a model or system behaviour made without notice to the institution. Contractually excluded; treated as an incident when discovered.

- **Subcontractor flow-down:** the contract sentence making every vendor obligation bind the vendor's subcontractors too.
- **Subgroup:** a part of the served population reported on separately (by age, sex, language, or the categories your law protects) so that an average cannot hide a group-level failure.
- **System Owner:** the named role accountable for a system's behaviour in service, senior enough to stop it without asking permission.
- **Tender (request for proposals):** the written document sent to vendors stating requirements, whose written answers can be made part of the contract.
- **Validation:** testing the system against your institution's own historical cases, scored against criteria fixed in advance.
- **Vendor:** the company selling the system; where a reseller sells it, the one who can answer the questions.

Annex C: References

All sources cited in this framework. Each was verified against the source during drafting.

Cases and studies

1. Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science*, volume 366, issue 6464 (25 October 2019), pages 447 to 453. doi:10.1126/science.aax2342.
2. Wong, A., Otlés, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrué, J., Phillips, M., Konye, J., Penzo, C., Ghous, M., and Singh, K. "External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients." *JAMA Internal Medicine*, volume 181, issue 8 (2021), pages 1065 to 1070. doi:10.1001/jamainternmed.2021.2626.
3. Habib, A. R., Lin, A. L., and Grant, R. W. "The Epic Sepsis Model Falls Short: The Importance of External Validation." *JAMA Internal Medicine*, volume 181, issue 8 (2021), pages 1040 to 1041. doi:10.1001/jamainternmed.2021.3333.
4. Ross, C. "Epic overhauls popular sepsis algorithm criticized for faulty alarms." *STAT News*, 3 October 2022.
5. Wong, A., Currey, D., Schwinne, M., Park-Egan, B., Meyer, S., Gutting, A., Cao, J., Khan, S., Dantes, R., Pan, T., Buchman, T., Singh, K., Bhavani, S. V., Lyons, P. G., Sjoding, M. W., and Tarabichi, Y. "Multicenter Prospective Validation of an Updated Proprietary Sepsis Prediction Model." *JAMA Network Open*, volume 9, issue 2 (2026), article e260181. doi:10.1001/jamanetworkopen.2026.0181.
6. Skitka, L. J., Mosier, K. L., and Burdick, M. "Accountability and automation bias." *International Journal of Human-Computer Studies*, volume 52, issue 4 (2000), pages 701 to 717. doi:10.1006/ijhc.1999.0349.
7. Parlementaire ondervragingscommissie Kinderopvangtoeslag, "Ongekend Onrecht" ("Unprecedented Injustice"), Netherlands House of Representatives, December 2020.
8. Autoriteit Persoonsgegevens (Dutch Data Protection Authority), "Boete Belastingdienst kinderopvangtoeslag," fine decision of 2.75 million euros, published 7 December 2021.
9. Amnesty International, "Xenophobic Machines: Discrimination through unregulated use of algorithms in the Dutch childcare benefits scandal," October 2021.
10. Information Commissioner's Office (United Kingdom), "Royal Free - Google DeepMind trial failed to comply with data protection law," news release and Data Protection Act 1998 undertaking, 3 July 2017.

Standards and instruments

11. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (the EU Artificial Intelligence Act). Entered into force 1 August 2024; obligations apply in phases. Amended by Regulation (EU) 2026/1744 of 8 July 2026 (Digital Omnibus on AI), in force 27 July 2026.
12. NIST, Artificial Intelligence Risk Management Framework (AI RMF 1.0), January 2023.
13. ISO/IEC 42001:2023, Information technology, Artificial intelligence, Management system. December 2023.
14. ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection, Information security management systems, Requirements.
15. OECD, Recommendation of the Council on Artificial Intelligence (the OECD AI Principles), adopted May 2019, updated 2024.
16. UNESCO, Recommendation on the Ethics of Artificial Intelligence, adopted by 193 member states, November 2021.
17. Regulation (EU) 2016/679 (the General Data Protection Regulation), cited as an example of a family of national data protection laws: Articles 22 and 33, Recital 26.

The limits of this framework, restated

This framework is guidance and it is not legal advice. It does not evaluate model internals; its method is external evidence, testing, monitoring, and contract. It certifies compliance with no law or standard; the mapping in Annex A shows direction of travel. It does not replace clinical, financial, or engineering judgement, and it does not govern the building of models. It names the points where paid specialist help is required: legal review before any high-stakes contract signature; local data protection advice before personal data crosses a border; independent security testing for public-facing systems; independent adversarial testing for systems where successful manipulation could affect safety, security screening, or access to essential services; and domain review before any high-stakes go-live. Where any national law or sector rule applies to you, it prevails over this framework.

About Yellow Sapphire Technologies

Yellow Sapphire Technologies Inc. (YS Tech) is a technology services company incorporated in New Brunswick, Canada, with its registered office in Moncton and offices in Dubai, Doha, London, and New Delhi. This framework was authored under the direction of Annuj Chauhan, Co-Founder and CEO. It was drafted by the team of analysts at Yellow Sapphire Technologies with the assistance of AI tools, under his direction and editorial control. He reviewed and approved the final text, and every citation in Annex C was checked against its published source before publication. In 2026, YS Tech was invited to contribute to the United Nations Global Dialogue on AI Governance and registered this framework as a public commitment after the Dialogue; this document is the delivery of that commitment.

How to cite this framework

Yellow Sapphire Technologies Inc. *Open AI Governance and Deployment Framework*, Version 1.0. Moncton, New Brunswick: Yellow Sapphire Technologies Inc., 1 October 2026. [doi:10.5281/zenodo.22986774](https://doi.org/10.5281/zenodo.22986774). Version 1.0 is deposited, on publication, in an open archival repository (Zenodo) under that DOI, so the reference is stable and permanent; each future version receives its own DOI, and citations should state the version.

Licence

This work is published under the **Creative Commons Attribution 4.0 International licence (CC BY 4.0)**. In practice this means you may copy, redistribute, adapt, translate, and build upon this framework, for any purpose including commercially, provided you give appropriate credit to Yellow Sapphire Technologies Inc. as the source, link to the licence, and indicate if changes were made. In one sentence: you are free to localise this document to your country, language, sector, and institution without asking permission, so long as you say where it came from. Two things the licence does not cover, because CC BY 4.0 does not license trademarks or names: the name “Open AI Governance and Deployment Framework,” the version designations, and the publisher’s names and marks are not licensed, and adaptations must be described as adapted from the Open AI Governance and Deployment Framework (stating the version adapted from) rather than presented as the framework itself or as endorsed by the publisher. Licence text: creativecommons.org/licenses/by/4.0.

Disclaimers

This framework is guidance and it is not legal advice. Institutions must consult qualified local counsel on their own legal obligations, including before signing contracts based on the clause library, before transferring personal data, and wherever this document says “check locally.”

On the United Nations. Publication of an initiative through the Partnerships Hub of the Global Dialogue on AI Governance does not constitute endorsement by the United Nations, the Co-Chairs, or the Global Dialogue. This framework is the sole work and responsibility of Yellow Sapphire Technologies Inc. The names and emblems of the United Nations are not used in this document as, and must never be presented as, a badge of approval.

On the cases cited. The documented cases in this framework are drawn from published rulings, peer-reviewed studies, official inquiries, and reporting, cited in Annex C. They are used to show why controls exist, and no statement in this framework should be read as a claim about any organisation beyond what the cited sources state.

Feedback and versioning

Feedback. Write to Anuj Chauhan, Co-Founder and CEO of Yellow Sapphire Technologies, at anuj@ystech.io when an artifact fails in practice, a checklist is missing a question, a passage does not survive translation, you find an error, or you have an improvement to suggest. The authors will read every message. Institutions adopting the framework can use the same address to arrange a conversation with the authors about implementation, at no charge. A contribution that leads to a material change in a checklist, clause, or definition is acknowledged in a later version, with the contributor's consent.

Versions and downloads. This is Version 1.0. Every file, including the editable artifacts, is free to download from ystech.io/ai-governance without registration. Revised versions are published there with a version history that records what changed and why. The publisher decides revisions. Earlier versions remain available, so each institution decides when to move to a new version; the system register (Artifact 7.11) records the version it adopted.

Translations. Community translations are welcome under the licence and will be linked from the framework page. A translation is official only when the publisher has commissioned and checked it, and every translation states the version it was made from.