Manifest AI

POLICY

Global AI Ethics Policy

How Manifest designs, builds and operates enterprise AI that is lawful, safe, fair, sovereign and under human control.



01
Ten Fair AI principles

02
Controls that scale with agent autonomy

03
Evidence every client can file

1. Policy statement and purpose

Manifest builds AI that acts inside enterprise and government systems of record, so every Manifest AI system must be lawful, safe, fair, sovereign and under human control for its whole life. This policy sets the rules that make that true, and it binds everyone who builds, sells, deploys or operates Manifest AI.

Enterprise AI has moved from answering questions to taking actions. A Manifest agent can read a contract, post a journal, raise a purchase order or answer a citizen. That changes the risk: a wrong answer can be ignored, but a wrong action can move money, change a record or affect a person's rights. This policy therefore goes further than general AI ethics statements in three ways:

1. **It governs actions, not only outputs.** Agents are classified by how much autonomy they have, and controls scale with that autonomy (section 5.3).
2. **It treats sovereignty as an ethical duty.** Where data, models and audit trails live is a decision for the client, not for Manifest or a model provider (principle 4.9).
3. **It produces evidence.** Every use case leaves a record a client can file with its own regulator, board or auditor (section 7.3).

The policy's objectives are to:

- protect the rights, safety and interests of the people Manifest AI affects, including people who never use it directly
- give clients the assurance and evidence they need to adopt AI with confidence
- meet Australian law and align with Australian and international AI frameworks (section 10)
- make responsible AI a working discipline in every engagement, not a statement on a website

Manifest's AI management system is certified to ISO/IEC 42001. Manifest puts this policy into practice through its Fair AI Assessment, delivery checklists and the evidence pack described in section 7.

2. Scope

This policy applies to every AI system Manifest designs, builds, sells, deploys, operates or uses, at every stage of its life, from idea to retirement.

2.1 People

- All Manifest employees, directors and contractors, including forward-deployed engineers and engineers in related entities who work on Manifest products
- Anyone who plays an active part in an AI system's life (an **Actor**): product, engineering, data, design, sales, delivery and operations
- Partners who resell, implement or extend Manifest solutions, as their agreements require

Every Actor must know which principles apply to their work, how the Fair AI Assessment works, and how to raise a concern.

2.2 AI systems

Type	Example	Covered
Manifest platform	Manifest aOS and its workspaces	Yes
Manifest agents	Pre-built Agent Hub agents, such as document processing and reconciliation	Yes
Client-specific AI	Custom agents and workflows built for one client, including client-owned IP	Yes
Proofs of value, pilots and trials	Time-limited builds on client data	Yes, with a lighter record for standard cases
Third-party models and services	Foundation models, embedding models, OCR and speech services inside a Manifest solution	Yes, through supplier controls (section 8)
AI tools Manifest uses internally	Coding assistants, meeting tools, sales tools	Yes, for data handling and red lines

2.3 Shared responsibility

Manifest, its clients and its model providers share responsibility for an AI system. Manifest is responsible for how it designs, builds and operates the system. The client is responsible for its use within the agreed purpose, its own business decisions and its own legal duties. Section 8.3 sets out who owns what.

This policy does not override the law. Where a requirement here conflicts with an applicable law or a client's binding policy, the stricter rule applies, and the conflict must be reported to the AI Governance Lead.

3. Key terms

Term	Meaning
AI system	A machine-based system that infers from its inputs how to produce outputs such as predictions, content, recommendations, decisions or actions. It includes its models, data pipelines, prompts, tools and interfaces.
AI use case	One application of an AI system to one business problem for one client or user group, for example invoice reconciliation for a mining client. Use cases, not platforms, are assessed.
AI model	The trained component that produces an output, such as a large language model, an embedding model or a classifier.
AI agent	An AI system that plans and takes actions through tools, for example reading a document, then updating a record in SAP or TechnologyOne.
Autonomy level	How far an agent can act without a person approving each action, rated A0 to A4 (section 5.3).
Principal	The person or role on whose authority an agent acts. An agent can never hold more rights than its principal.
System of record	A system whose data is relied on as the official record, such as an ERP, finance, HR, asset or case management system.
Grounding	Basing an output on identified source documents or data, with a link back to the source.
Affected person	Anyone whose rights, money, services, safety or job can be changed by an AI system's output, whether or not they use it.
Automated decision	A decision with a legal or similarly significant effect on a person, made without a meaningful human check before it takes effect.
Human oversight	Human-in-the-loop (a person approves each output or action), human-on-the-loop (a person monitors and can step in) or human-in-command (a person can override, pause or shut down the system at any time).
Sovereign deployment	Running data, models and audit trails inside a jurisdiction and environment the client controls, up to fully on-premises and air-gapped.
Evidence pack	The records of a use case's assessment, tests and sign-off that Manifest gives the client.
AI incident	An event where an AI system causes, or nearly causes, harm, a data breach, an unlawful outcome, an unsafe action or a loss of human control.

4. Fair AI principles

Manifest's ten principles apply to every use case together; no principle can be traded away to meet another. Principles 1 to 8 follow Australia's AI Ethics Principles. Principles 9 and 10 are Manifest's own.



Principles 9 and 10, shaded, are Manifest's own.

Requirements that apply under every principle

- Every use case passes the Fair AI Assessment before design starts, and again on any material change.
- Data used to build, ground or test a system must be accurate, relevant and fit for the use case.
- Every system logs its inputs, outputs and actions so that any outcome can be traced, within privacy limits.
- Every system is monitored in production for accuracy, drift, misuse and security events.

4.1 Wellbeing and proportionality

AI must deliver a clear benefit and must not do more than its purpose needs.

- Each use case states its purpose, its intended users, its out-of-scope uses and the measure of success, before build.
- The team must show why AI is needed, and why a rule, a report or a simpler model would not do.
- The smallest model and the lowest autonomy level that meet the purpose must be used.

4.2 Human-centred values

AI must respect human rights and be designed around the people who use it and the people it affects.

- Affected people are identified in discovery, not only users.
- Users must be able to work without the AI where their role allows, and must not be pressured to accept its output.
- No system may be designed to create emotional dependence or to pretend to be human.

4.3 Fairness

AI must not discriminate unlawfully or produce unjustly unequal outcomes.

- Each use case defines what fair means in its context, and how it will be measured, before build.
- Systems that affect people are tested for bias before go-live, including the output of third-party models whose training data is unknown.
- Fairness is monitored in production, and affected people can challenge an outcome they believe is unfair.
- Interfaces meet accessibility standards (WCAG 2.2 AA as the target).

4.4 Privacy and security

AI must protect personal information and resist attack.

- Personal information is collected, used, kept and deleted in line with the Privacy Act 1988 and the Australian Privacy Principles, and the client's own obligations.
- Only the personal information the purpose needs is used. Prompts, logs and agent memory are treated as personal information where they contain it.
- Client data is never used to train or fine-tune models for anyone other than that client, and fine-tuning for the client needs its written consent.
- Systems follow Manifest's certified controls (ISO/IEC 42001 for AI management, SOC 2 and ISO 27001 for security), including threat modelling, red-teaming and defences against prompt injection and data poisoning.

4.5 Reliability and safety

AI must work as intended, within known limits, and fail safely.

- Acceptable error rates are agreed with the client before build, and tested against real scenarios.
- Generative output in a decision-support or official context must be grounded in identified sources.
- Every system has a fallback: when confidence is low or a control fails, it hands the case to a person or stops.
- Agent actions must be reversible where possible, and irreversible actions need human approval (section 6.3).

4.6 Transparency and explainability

People must know when they deal with AI, and be able to understand its output.

- Users are told, in the interface, when content or an action comes from AI.
- Outputs show their sources, and show when they are not grounded.
- On request, an affected person can get an explanation of how an output that affected them was produced.
- Each use case keeps a model card and a data card describing purpose, limits, data, models and tests.

4.7 Contestability

People affected by an AI-supported outcome must be able to challenge it.

- Every system that affects people has a route for a person to ask for human review.
- Where an automated decision is used, the client's review route is designed and tested before go-live.
- Upheld challenges feed back into testing and design.

4.8 Accountability

People, never AI, are accountable for AI outcomes.

- Every use case has a named owner at Manifest and a named risk owner at the client.
- Every agent acts on the authority of a named principal, and every action is traceable to it.
- Decision rights for approving, changing and retiring a system are written down (section 7).

4.9 Data and model sovereignty

The client decides where its data, models and audit trails live.

- The deployment option (Manifest cloud, client private cloud, or on-premises including air-gapped) and data residency are agreed in writing before build and tested before go-live.
- No data, prompt or log may leave the agreed jurisdiction or go to a new provider without the client's written approval.
- The platform stays model-agnostic, so a client is never locked to one model provider.

4.10 Respect for communities and environment

AI must respect the communities it serves and limit its environmental cost.

- Systems that serve a community are shaped with that community's input where practical.
- Data about First Nations peoples is handled under Indigenous data sovereignty principles, with consent from the right authority.
- Teams choose efficient models, route simple tasks to smaller models and avoid AI where it adds no value.

5. Risk classification

Every use case gets two ratings at intake: a risk tier (prohibited, high-risk or standard) and an autonomy level (A0 to A4). Together they set the controls, the approvals and how often the use case is re-assessed.

5.1 Prohibited use cases (red lines)

Manifest will not design, build, sell, deploy or operate an AI system, for any client or fee, that:

Personal freedom

- carries out surveillance that targets people or groups, including by biometrics, facial recognition or emotion recognition
- infers emotions in a workplace or place of education, except for medical or safety reasons
- scores or classifies people on their social behaviour or personal traits in ways that lead to unjust treatment
- re-identifies people in data that was anonymised or de-identified

Harm to people and society

- is designed to deceive or manipulate people against their own interests, including by exploiting age, disability or financial hardship
- undermines public debate, elections or democratic processes
- discriminates against people or denies a group fair access to services or opportunities
- is used to design, target or operate weapons

Sovereignty and law

- moves client or citizen data, or models trained on it, out of the agreed jurisdiction without the client's written approval
- breaches the Privacy Act 1988, anti-discrimination law, the client's own law, or a prohibited practice under Article 5 of the EU AI Act

Loss of control

- lets an agent grant itself new permissions, change its own limits, or switch off its own logging, monitoring or oversight

If an Actor finds that a use case crosses a red line at any stage, work stops and the AI Governance Lead is told the same day.

5.2 High-risk use cases

A use case is high-risk if any one of these applies:

1. It processes personal information in training, prompts, outputs or agent memory.
2. Its outputs can affect a person's rights, services, money, safety, health or job.
3. It makes or drives an automated decision about a person.
4. It is used in government services, critical infrastructure, health, finance, insurance, employment, education, law enforcement or migration.
5. It is an agent at autonomy level A2 or above that acts in a system of record.
6. Data, prompts or a model leave the client's environment or jurisdiction, or go to a third-party provider.

High-risk use cases are permitted, but only after the full review in section 7.2. Most public-sector and financial-services work will be high-risk; that is expected.

All other use cases are **standard**. They proceed under normal delivery controls with a lighter assessment record.

5.3 Agent autonomy levels

The more an agent can do without a person, the stronger the controls must be.

Level	What the agent can do	Human oversight	Example
A0 Inform	Search, summarise and answer; takes no action	User judges the output	Policy question answering with citations
A1 Draft	Prepares a draft, entry or recommendation that a person must submit	Human-in-the-loop by design	Drafts a variation claim for a contract manager
A2 Act with approval	Carries out an action after a person approves that action	Human-in-the-loop for each action	Posts a journal once a finance officer approves it
A3 Act within limits	Acts alone inside pre-set limits on value, scope and reversibility; sends exceptions to a person	Human-on-the-loop, with sampling and alerts	Auto-matches invoices under an agreed value and tolerance
A4 Open autonomy	Acts across open-ended tasks with no pre-set limits	None in real time	Not permitted in client production

5.4 What each combination needs

Risk tier	Allowed autonomy	Approval to go live	Re-assessment
Standard	A0 to A3	Use Case Owner, with a record checked by the AI Governance Lead	On any material change, and at least every 12 months
High-risk	A0 to A2; A3 only with Review Panel approval and hard limits	AI Ethics Review Panel and the client risk owner	On any material change, and at least every 6 months
High-risk with an automated decision about a person	A0 to A2 only; a person approves every decision	AI Ethics Review Panel, client risk owner, and Board notified	On any material change, and at least every 6 months
Prohibited	None	Not permitted	Not applicable

A material change includes a new model or major model version, a new data source, a new tool or system the agent can act in, a higher autonomy level, a new user group or a new jurisdiction.

6. Requirements for enterprise and agentic AI

These controls turn the principles into engineering rules for the systems Manifest actually builds: agents that read enterprise data and act in systems of record.

6.1 Data

- Every data source is recorded in the data card with its owner, purpose, sensitivity, residency and retention.
- Agents read data with the access rights of their principal, never a broader service account.
- Document lineage is kept, so every grounded answer and every agent action can be traced to its source records.
- Data is deleted or returned at the end of an engagement as the contract requires, including copies in vector indexes, caches, logs and agent memory.

6.2 Models

- Manifest keeps a model register: every model in use, its provider, version, hosting location, data-retention terms and approved use cases.
- A model is chosen for each use case on tested accuracy, bias, toxicity, latency, cost and where it can run, and the results go in the model card.
- A new model or major version is a material change: it is re-tested on the use case's evaluation set before release.
- Third-party model providers must not store client data beyond processing or use it for training (section 8.2).

6.3 Agents

- **Identity.** Each agent has its own identity, separate from any person, so its actions are distinguishable in every log.
- **Least privilege.** An agent gets only the tools, systems and permissions its task needs, and never more than its principal holds.
- **Hard limits.** A3 agents run inside written limits on transaction value, record types, volume per period and the systems they can touch. The limits are enforced by the platform, not by the prompt.
- **Separation of duties.** An agent cannot approve its own action, and two agents cannot approve each other's.
- **Irreversible actions.** Payments, deletions, external messages and changes to a person's entitlements need human approval, whatever the autonomy level.
- **Rollback.** Agent actions are reversible where the target system allows, and the rollback path is tested before go-live.
- **Stop control.** Every agent can be paused or shut down at once by the Use Case Owner and the client, without a code change.
- **Memory.** What an agent remembers between sessions is limited to its purpose, visible to the client, and covered by the same retention and deletion rules as other data.
- **Multi-agent systems.** When agents hand work to other agents, the chain of authority back to the principal is kept and logged.

6.4 Generative AI

- Outputs used for decisions or official records are grounded in client sources and show citations; ungrounded content is labelled as such.

- Content filters screen inputs and outputs for harmful, unlawful and off-purpose content.
- AI-generated documents, messages and images are marked as AI-generated, in a machine-readable form where practical.
- Prompts and system instructions are version-controlled and change-managed like code.

6.5 Security

- Everything an agent reads, including documents, emails, web pages and tool results, is treated as untrusted input that may carry hidden instructions (prompt injection).
- Secrets and credentials are never placed in prompts, and agents reach them only through a managed vault.
- High-risk systems are red-teamed before go-live and after any material change, covering jailbreaks, prompt injection, data leakage and misuse of tools.

6.6 Monitoring

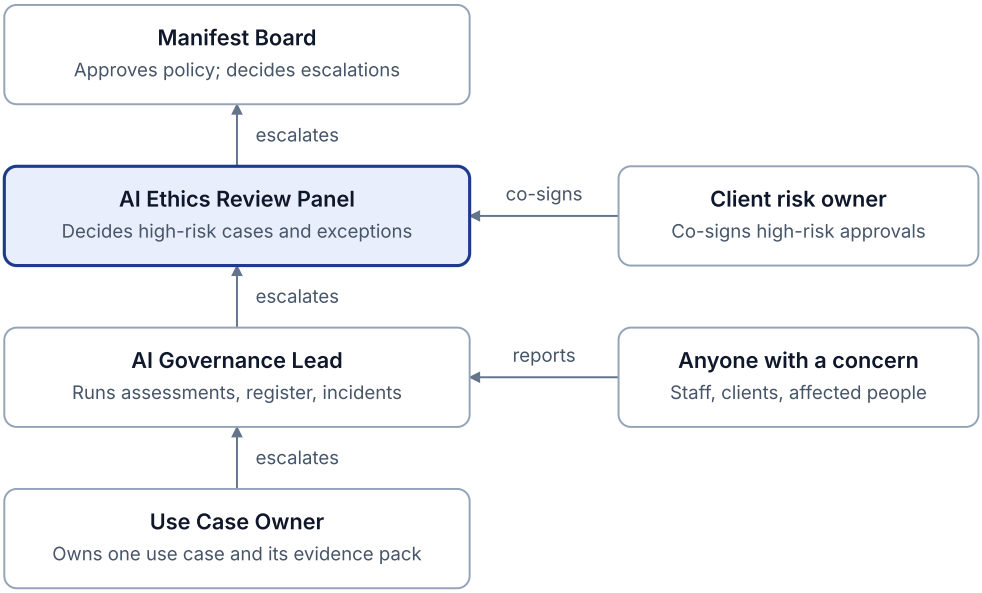
Each production system has agreed thresholds and alerts for:

- accuracy against the go-live baseline, and drift from it
- fairness measures for the groups identified at design
- the share of cases sent to a person, and how often people override the AI
- blocked actions, limit breaches, security events and user complaints

A breach of a threshold is handled as an incident under section 9.

7. Governance

The Board owns this policy, the AI Ethics Review Panel decides high-risk cases, and the AI Governance Lead runs the process day to day.



Decisions go up one level when the level below cannot resolve a risk, or when this policy requires it.

7.1 Governance bodies and roles

Body or role	Members	Decision rights and duties
Manifest Board	Board directors	Approves this policy and its annual review. Has final accountability for high-risk use cases. Decides escalated cases. Receives a quarterly AI assurance report.
AI Ethics Review Panel	CEO, CTO, AI Governance Lead; privacy, legal and security advisers as needed	Approves, approves with controls, or declines high-risk use cases. Approves A3 autonomy for high-risk cases. Grants or refuses exceptions. Can reclassify a case.
AI Governance Lead	Appointed by the CEO	Owns the Fair AI Assessment, the AI register and the model register. Screens every intake. Manages incidents. Maintains this policy and its supporting procedures. First contact for questions and concerns.
Use Case Owner	Manifest engagement lead	Accountable for one use case end to end. Completes the assessment and evidence pack. Ensures every control is in place.
Actors	Engineers, data scientists, designers, delivery and sales staff	Build, sell and operate to this policy. Record what they build and test. Raise concerns.
Client risk owner	Named by the client	Co-signs high-risk approvals and owns the system in the client's risk register.
External advisers	Independent experts, invited as needed	Give independent views on novel or contested cases.

7.2 The Fair AI Assessment

Every use case goes through the Fair AI Assessment before design starts:

1. **Intake.** The Use Case Owner describes the system, data, users, affected people and context.
2. **Red-line check.** Any red line stops the use case.
3. **Classification.** The risk tier and autonomy level are set (section 5).
4. **Review.** Standard cases: the AI Governance Lead checks the record. High-risk cases: a joint values assessment with the client, a controls and test plan, and a Review Panel decision.
5. **Sign-off.** Go-live needs the approvals in section 5.4 and a complete evidence pack.
6. **Ongoing assurance.** Monitoring, incident handling and re-assessment on schedule or on change.

7.3 Records and the AI register

- Manifest keeps an AI register of every use case, with its tier, autonomy level, owners, models, deployment and next review date.
- Every use case keeps an evidence pack: intake, classification, values assessment (high-risk), data card, model card, test and red-team results, oversight design and sign-off.
- Clients receive their evidence pack and an entry for their own AI register, in a form that maps to the Queensland FAIRA framework and the national Guidance for AI Adoption.

- Records are kept for at least seven years after a system is retired, or longer if the client's rules require it.

7.4 Training

- Every Actor completes responsible AI training before working on a use case, and refreshes it every year.
- Engineers working on agents complete additional training on agent security, prompt injection and oversight design.

8. Partners, suppliers and third-party models

Manifest's duties do not stop at its own code: partners and suppliers who touch a Manifest AI system must meet the same standard, and contracts must say so.

8.1 Partners

- Partners who resell, implement or extend Manifest solutions must follow this policy for that work, and their agreements must require it.
- Use cases a partner builds on Manifest aOS go through the Fair AI Assessment like any Manifest use case.
- Partners pass any concern or incident to the AI Governance Lead within 24 hours.

8.2 Model providers and other suppliers

Before a third-party model or AI service is used with client data, Manifest confirms in writing that the supplier:

- processes client data only to provide the service, and does not use it to train or improve its models
- does not keep prompts or outputs beyond what the service needs, or keeps them only for an agreed abuse-monitoring period
- hosts and processes data in the jurisdictions the client has approved
- meets security standards at least equal to Manifest's, and reports breaches promptly
- publishes the model's intended uses, limits and known risks

Suppliers that cannot meet these terms may be used only for standard use cases with no personal or confidential data, or not at all.

8.3 Shared responsibility

Area	Manifest	Client	Model provider
Purpose and lawful basis for use	Advises; refuses red lines	Decides and owns	Not involved
Fair AI Assessment and evidence pack	Runs and produces	Co-signs high-risk cases	Supplies model information
Data quality and access rights	Designs controls and least-privilege access	Owns the data and grants access	Not applicable
Model choice and testing	Selects, tests and records	Approves for high-risk cases	Publishes model limits and safety testing
Deployment and data residency	Builds to the agreed option	Chooses the option	Honours agreed region and retention
Human oversight in daily use	Designs checkpoints and stop controls	Staffs the checkpoints and uses them	Not applicable
Monitoring and incidents	Monitors, alerts, investigates	Reports issues; acts on its side	Reports model incidents and changes
Decisions about people	Never makes them alone	Accountable for every decision	Not applicable

9. Concerns, incidents, compliance and exceptions

Anyone can raise a concern at any time, every AI incident is acted on by severity, and a use case found to breach this policy is changed, paused or withdrawn.

9.1 Raising a concern

- Staff and partners raise concerns with their Use Case Owner, the AI Governance Lead, or through Manifest's internal reporting channel.
- Clients, users and affected people can raise concerns with Manifest through manifestai.com.au/contact, or through the client. Concerns can be raised anonymously.
- No one will suffer any detriment for raising a concern in good faith. Concerns that qualify are also protected under the whistleblower provisions of the Corporations Act 2001.

9.2 AI incidents

Severity	Examples	First action	Tell the AI Governance Lead	Tell the client
Critical	Harm to a person; red line crossed; personal data exposed; agent acts outside its limits or cannot be stopped	Pause or shut down the system at once	Within 1 hour	Within 24 hours, or sooner if the contract says so
High	Biased or unlawful outcomes; repeated wrong actions in a system of record; successful prompt injection	Lower autonomy or add human approval until fixed	Within 24 hours	Within 2 business days
Moderate	Monitoring threshold breached; ungrounded output in an official context; upheld complaint	Investigate and correct	Within 5 business days	In the next service report

For each incident the AI Governance Lead:

1. confirms the facts and the severity, and makes sure the system is safe
2. refers critical incidents to the Review Panel, which decides on corrective action and escalates to the Board where needed
3. records the cause, the fix and the lessons, and updates tests and controls
4. reports all incidents to the Board each quarter

Where an incident involves personal information, Manifest supports the client's obligations under the Notifiable Data Breaches scheme in the Privacy Act 1988, and meets its own.

9.3 Corrective measures

If a use case in production is found to cross a red line or to breach this policy, Manifest will adapt it, restrict it, or withdraw it, in agreement with the client where the contract allows. Manifest will not keep operating a system that crosses a red line.

9.4 Breaches of this policy

A breach of this policy by an employee or contractor may lead to disciplinary action, up to ending employment or the contract. A breach by a partner or supplier may lead to the agreement being suspended or ended.

9.5 Exceptions

- Exceptions to this policy must be requested in writing to the AI Governance Lead, with the reason, the risk and the compensating controls.
- Only the AI Ethics Review Panel can grant an exception. It must be time-limited, recorded in the AI register, and reported to the Board.
- No exception can be granted to a red line.

10. Alignment, review and related documents

This policy is built so that a client can map Manifest's controls straight onto the framework it already answers to.

10.1 Alignment to laws and frameworks

Law or framework	Where this policy meets it
Australia's AI Ethics Principles	Principles 4.1 to 4.8
Guidance for AI Adoption (AI6), National AI Centre	Accountability (7), impacts (5, 7.2), risk management (5, 6), information sharing (4.6, 7.3), testing and monitoring (6.6), human control (5.3, 6.3)
Policy for the responsible use of AI in government, Digital Transformation Agency	Evidence and transparency material that helps Commonwealth agency clients meet their obligations (7.3)
FAIRA framework, Queensland Government	Intake and values assessment in the evidence pack (7.2, 7.3)
Privacy Act 1988 and the Australian Privacy Principles	Principle 4.4; section 9.2
ISO/IEC 42001 (AI management systems)	Manifest is certified to ISO/IEC 42001. This policy is a core part of its certified AI management system (sections 5 to 9).
NIST AI Risk Management Framework	Govern, map, measure and manage functions across sections 5 to 9
EU AI Act, Article 5	Red lines in section 5.1

10.2 Review

- The AI Governance Lead reviews this policy at least once a year, and sooner when the law, national guidance or Manifest's products change in a way that affects it.
- The Board approves every new version.

10.3 Related documents

- Manifest AI management system (ISO/IEC 42001 certified) and information security policies (SOC 2 and ISO 27001 controls)
- Manifest privacy policy
- Manifest partner and supplier agreements
- Manifest Trust Centre