

Responsible AI Fundamentals

- Build working knowledge for AI project teams, product managers, and educators
- Define responsible AI, recognize key risks, and apply core principles
- Take practical next steps toward safe, ethical, and trustworthy AI
- Roadmap: principles, risks, lifecycle roles, governance, tools, teaming, and action



What Responsible AI Means



- Develop, deploy, and operate AI safely, ethically, and legally



- Reduce bias and harm while earning user trust



- Comply with emerging regulations and protect reputation



- A shared technical and organizational discipline



- Trust must be centered in every AI system lifecycle



Why Responsible AI Matters

- Real-world failures show concrete harm: defamation, plagiarism, unsafe chatbot behavior
- Legal, business, and reputational consequences now include court liability and fines
- Trust and user adoption are strategic advantages
- Teams that address risk early avoid costly retrofits and maintain credibility



Core Principles

- ▶ Fairness, accountability, and transparency form the shared baseline
- ▶ Privacy, safety, robustness, and human oversight complete the set
- ▶ OECD AI Principles and NIST AI RMF offer practical guidance
- ▶ Principles gain meaning through design choices, evaluation, and documentation
- ▶ Team practices turn principles from words into reviewable actions



Types of Risk and Harm



- Bias, privacy violations, misinformation, and downstream societal harms



- AI-specific harms often bypass traditional security frameworks



- Misleading ads, defamatory content, and impersonation damage trust



- Early warnings: prompt injection, data leakage, unusual agent behavior



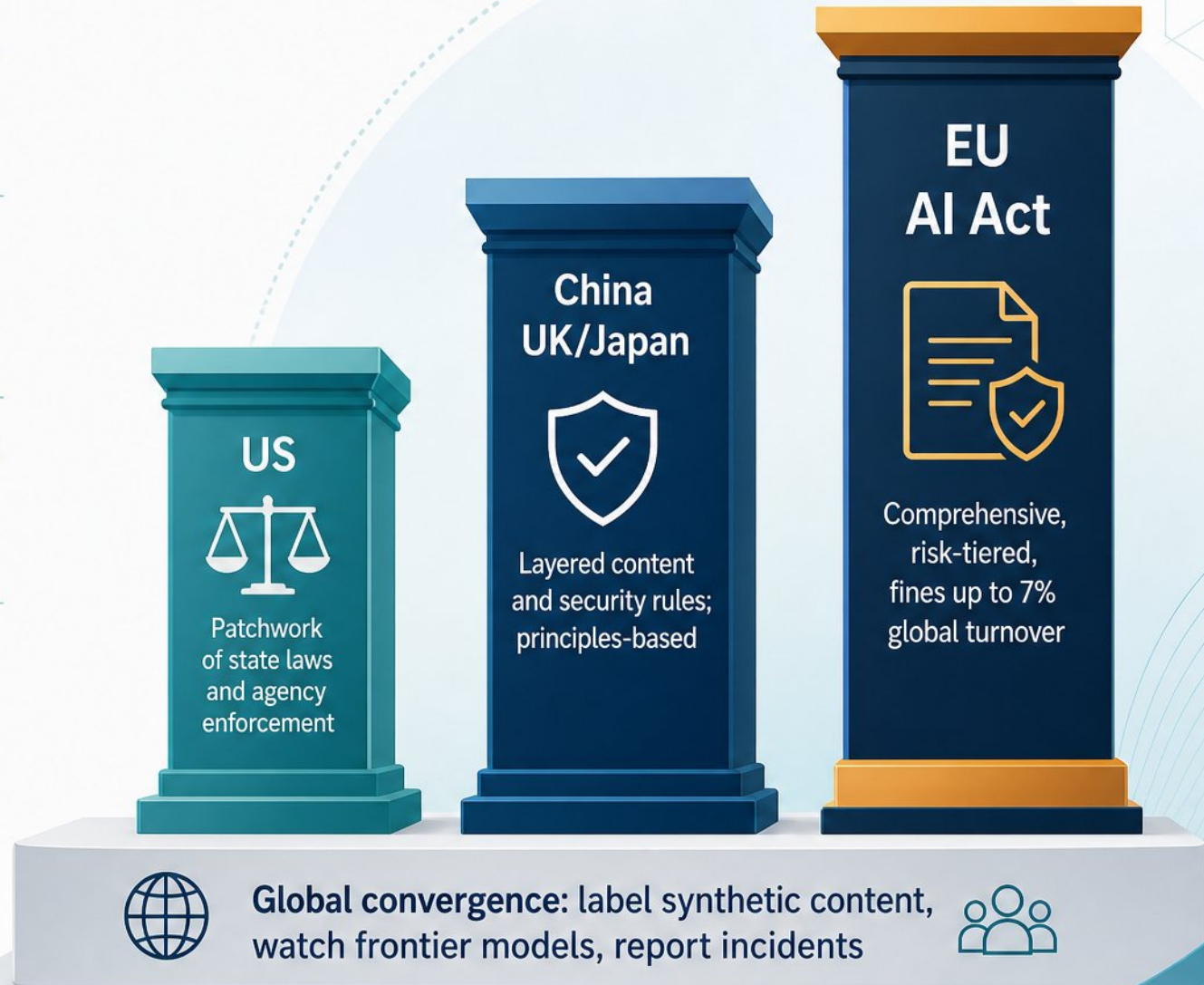
Understanding the NIST AI RMF

- **Govern, Map, Measure, Manage:** four core functions
- **Govern** is cross-cutting: policies, accountability, risk culture
- **Map** establishes context for use and affected people
- **Measure** and **Manage** drive assessment, response, monitoring



Regulatory Landscape

- EU AI Act: comprehensive, risk-tiered, fines up to 7% global turnover
- US: no federal law; patchwork of state laws and agency enforcement
- China: layered content and security rules; UK/Japan: principles-based
- Global convergence: label synthetic content, watch frontier models, report incidents



Lifecycle Responsibilities



Responsible duties span data, model, deploy, operate, decommission



Product managers own intended-use boundaries



Engineers own technical safeguards



Team leads own accountability and escalation



Document handoffs: assumptions, limitations, approvals



Governance and Policy

- Policies, review boards, and approval workflows create accountability
- Document everything: audit trails and model cards enable evidence-based review
- Maintain approved tool lists and data handling rules
- Define human review requirements before deployment
- Prepare incident response procedures in advance



Risk Management in Practice



- Assess risk continuously across operational, legal, data, and ethical dimensions



- Avoid scaling AI without governance, strong data quality, and human oversight



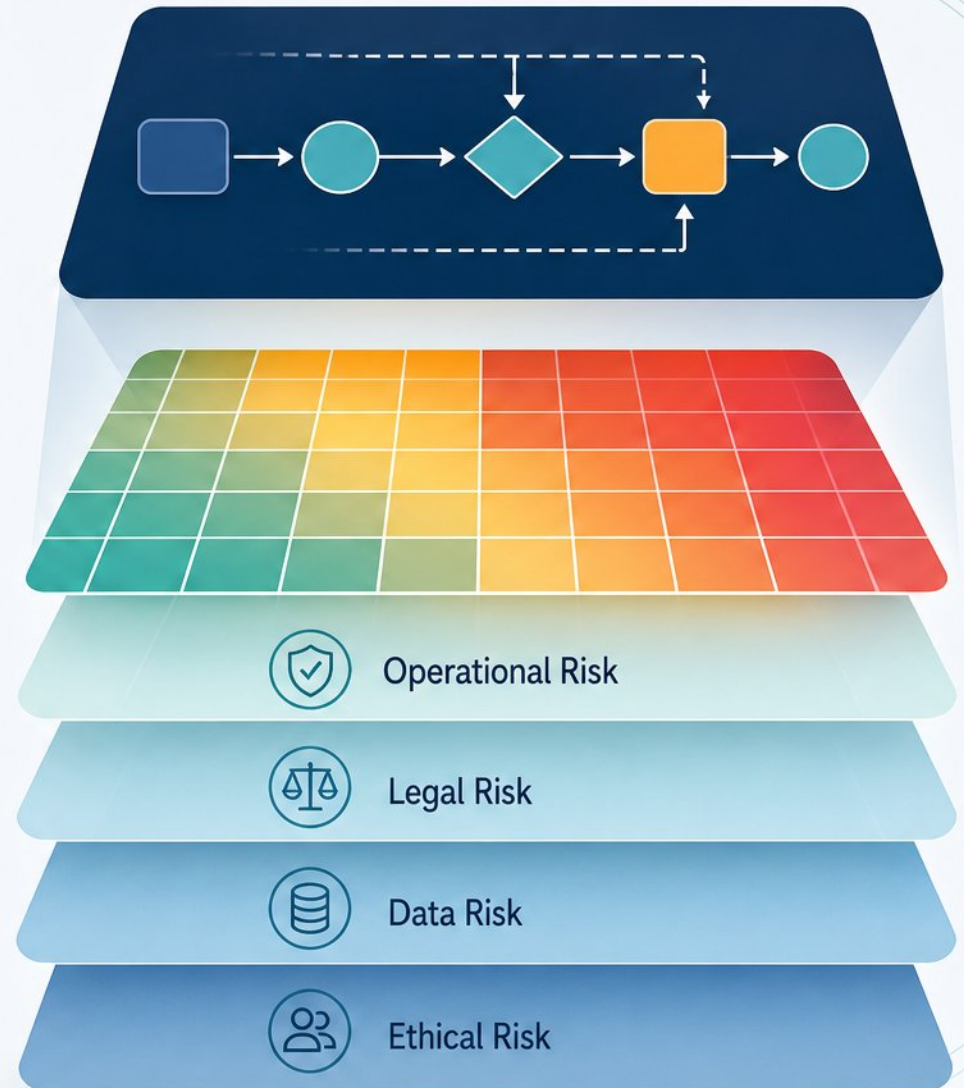
- Monitor for shadow AI and agent activity outside approved channels



- Document AI asset inventory and assign executive ownership



- Establish risk tolerance per use case before broadening deployment



Practical Tools and Methods



- Bias detection, fairness metrics, and generative AI safety tests



- Model cards, data cards, risk files, and implementation checklists for audit-ready docs



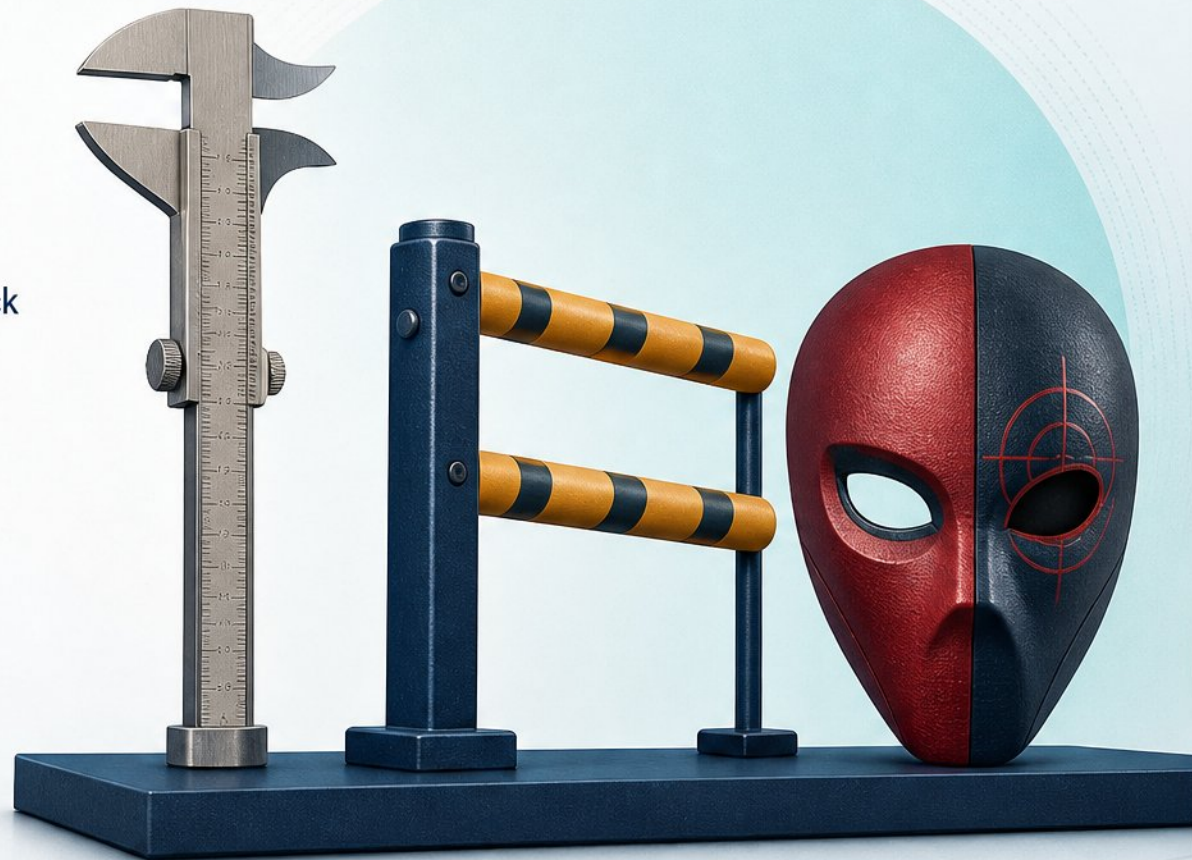
- Human review, red teaming, feedback loops, and guardrails balance automation with oversight



- Open-source toolkits: guardrails, prompt-injection firewalls, and compliance pipelines



- FRAI CLI and ASSERT support evidence-backed evaluation and documentation



Building Responsible Teams



- Diverse perspectives and domain experts surface risks technical teams miss



- Psychological safety makes early escalation the default



- Role clarity, training, incentives, and ongoing review beat one-time policies



- Senior leadership owns AI oversight, not just technology functions



- Innovation requires empowerment paired with stronger governance



Measuring, Reporting, and Responding



- Track model quality, drift, bias, and safety incidents



- Build AI-specific playbooks with staged containment



- Prepare communication channels before incidents occur



- Monitor continuously; AI behavior is non-deterministic



Action Plan and Next Steps



- Start with high-impact practices: define use cases, establish ownership, document governance



- 30-day checklist: inventory AI systems, identify data sensitivity, stand up basic review



- 60-day checklist: add evaluation, monitoring, and incident response for highest-risk systems



- 90-day checklist: implement documentation standards, expand training, formalize continuous review



- Reliable resources: NIST AI RMF, OECD AI Principles, EU AI Act guidance, open-source toolkits



PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/0df12b6f/public