AI PRODUCT MANAGEMENT 2 · DETAILED COURSE

Responsible AI, Safety and Governance

Managing the ethical, legal and security risks of AI products so they can be launched and trusted.

Chapter 18 of 26 · 24 topics · Edition 2026 · skillnavora.in

CHAPTER 18

Responsible AI, Safety and Governance

Managing the ethical, legal and security risks of AI products so they can be launched and trusted.

Responsible AI Fundamentals

Principles of fairness, reliability, safety, privacy, transparency, accountability and inclusiveness.

Responsible AI principles include fairness, reliability, safety, privacy, transparency, accountability and inclusiveness. Turn each principle into a testable requirement for your product.

AI Ethics

Weighing benefits and harms to people and society when designing and deploying AI.

Ethics weighs the benefits and harms to people and society when designing and deploying AI. Ask who could be harmed and who has not been consulted.

Data Privacy

Respecting what personal data is collected, how it is used and who sees it.

Respect what personal data is collected, how it is used and who can see it. Collect only what the feature needs and explain it clearly to users.

Data Protection

Technical and organisational safeguards such as encryption, access control, retention limits and minimisation.

Safeguards include encryption, access control, retention limits and data minimisation. Retention limits mean deleting data when it is no longer needed.

GDPR

The EU regulation giving people rights over personal data and imposing duties on organisations processing it.

The EU regulation gives people rights over their personal data and places duties on organisations that process it. It applies to products serving EU residents, even when the company is based elsewhere.

India's Digital Personal Data Protection Act

India's 2023 law on processing digital personal data, based on consent and notice, with obligations for data fiduciaries; confirm current rules with counsel.

India's 2023 law regulates digital personal data, based on consent, notice and obligations for data fiduciaries. Check current rules and their implementation timeline with legal counsel.

AI Security

Protecting models, data, prompts and tools from misuse and attack.

Protect models, data, prompts and tools from misuse and attack. Security reviews should cover the model provider, the data flow and the tools the model can call.

Data Leakage

Sensitive information exposed through prompts, outputs, logs or training data.

Sensitive information can leak through prompts, outputs, logs or training data. Redact personal data before it reaches a model or a log, when the feature does not need it.

Prompt Injection

Malicious text that tries to override instructions or trigger unwanted actions, including through retrieved content.

Malicious text tries to override instructions or trigger unwanted actions, including text hidden in retrieved documents. Separate instructions from data and limit the tools available to the model.

Jailbreak Risks

Prompts designed to bypass safety rules; test regularly and layer defences.

Jailbreak prompts try to bypass safety rules. Test regularly with known patterns, and layer defences rather than relying on one filter.

AI Bias

Systematic unfairness from skewed data, labels or design.

Bias is systematic unfairness from skewed data, labels or design choices. Audit outcomes across groups, not only overall accuracy.

Fairness

Ensuring outcomes are equitable across groups, measured with agreed metrics.

Fairness means outcomes are equitable across groups, measured with agreed metrics. There is no single fairness metric, so choose the one that matches the harm you are guarding against.

Explainability

Giving understandable reasons for outputs, especially for high-stakes decisions.

Explainability gives understandable reasons for outputs, especially for high-stakes decisions such as credit. Users and regulators may need the reasons, not only the score.

Transparency

Being clear about when AI is used, its limits and how data is handled.

Be clear about when AI is used, its limits and how data is handled. Users calibrate trust better when they know what the system is.

Human Oversight

Meaningful human review, override and accountability for important decisions.

Meaningful human review, override and accountability for important decisions. A human who rubber-stamps every output does not provide real oversight.

AI Guardrails

Input and output filters, policy checks and permission limits that constrain behaviour.

Input and output filters, policy checks and permission limits constrain behaviour. Guardrails work best as several layers rather than one rule.

Content Moderation

Detecting and handling harmful or disallowed content from users and models.

Detect and handle harmful or disallowed content from users and models. Define the policy first, then build tools to enforce it, with an appeals path.

Access Control

Limiting who and what can reach data, tools and models, based on role and need.

Limit who and what can reach data, tools and models, based on role and need. Apply the same rules to the AI feature as to the underlying data.

AI Governance

Policies, roles, review processes and standards for responsible AI across the organisation.

Governance sets policies, roles, review processes and standards for responsible AI across the organisation. It answers who approves an AI use case and who is accountable when it fails.

AI Risk Management

Identifying, rating and mitigating AI risks, using frameworks such as NIST AI RMF.

Identify, rate and mitigate AI risks. Frameworks such as the NIST AI Risk Management Framework give a structure for this work.

Enterprise AI Governance

Central oversight of AI use cases, vendors, data and approvals across a large organisation.

Central oversight of AI use cases, vendors, data and approvals in large organisations. It prevents teams from adopting unapproved tools with customer data.

Regulatory Compliance

Meeting laws and standards that apply to your market and use case, such as the EU AI Act, GDPR and sector rules.

Meet the laws and standards that apply to your market and use case, such as the EU AI Act, GDPR and sector rules. Compliance requirements shape product design, so involve legal early.

AI Incident Management

Detecting, triaging, fixing and learning from AI failures, with communication and post-incident review.

Detect, triage, fix and learn from AI failures, with clear communication and a post-incident review. Blameless reviews find systemic causes.

Responsible AI Product Launch

A checklist before launch covering risks, testing, approvals, user disclosure, monitoring and rollback.

A pre-launch checklist covers risks, testing, approvals, user disclosure, monitoring and rollback. Launch only when every item has an owner and a status.

Summary of terms

Term	Meaning
Responsible AI Fundamentals	Principles of fairness, reliability, safety, privacy, transparency, accountability and inclusiveness.
AI Ethics	Weighing benefits and harms to people and society when designing and deploying AI.
Data Privacy	Respecting what personal data is collected, how it is used and who sees it.
Data Protection	Technical and organisational safeguards such as encryption, access control, retention limits and minimisation.
GDPR	The EU regulation giving people rights over personal data and imposing duties on organisations processing it.
India's Digital Personal Data Protection Act	India's 2023 law on processing digital personal data, based on consent and notice, with obligations for data fiduciaries; confirm current rules with counsel.
AI Security	Protecting models, data, prompts and tools from misuse and attack.
Data Leakage	Sensitive information exposed through prompts, outputs, logs or training data.
Prompt Injection	Malicious text that tries to override instructions or trigger unwanted actions, including through retrieved content.
Jailbreak Risks	Prompts designed to bypass safety rules; test regularly and layer defences.
AI Bias	Systematic unfairness from skewed data, labels or design.
Fairness	Ensuring outcomes are equitable across groups, measured with agreed metrics.
Explainability	Giving understandable reasons for outputs, especially for high-stakes decisions.
Transparency	Being clear about when AI is used, its limits and how data is handled.
Human Oversight	Meaningful human review, override and accountability for important decisions.
AI Guardrails	Input and output filters, policy checks and permission limits that constrain behaviour.
Content Moderation	Detecting and handling harmful or disallowed content from users and models.
Access Control	Limiting who and what can reach data, tools and models, based on role and need.

AI Governance	Policies, roles, review processes and standards for responsible AI across the organisation.
AI Risk Management	Identifying, rating and mitigating AI risks, using frameworks such as NIST AI RMF.
Enterprise AI Governance	Central oversight of AI use cases, vendors, data and approvals across a large organisation.
Regulatory Compliance	Meeting laws and standards that apply to your market and use case, such as the EU AI Act, GDPR and sector rules.
AI Incident Management	Detecting, triaging, fixing and learning from AI failures, with communication and post-incident review.
Responsible AI Product Launch	A checklist before launch covering risks, testing, approvals, user disclosure, monitoring and rollback.

Practice exercise: Run a risk review for one AI feature covering privacy, bias, injection and misuse. Assign an owner and mitigation to each risk.

Self-check

1. Explain Responsible AI Fundamentals in two sentences to a non-technical stakeholder, then say what you would measure for it.
2. Explain Explainability in two sentences to a non-technical stakeholder, then say what you would measure for it.
3. Explain Responsible AI Product Launch in two sentences to a non-technical stakeholder, then say what you would measure for it.

Related SkillNavora material

- AI Governance and Responsible AI : Risk, safety and compliance.
- AI Tutor : Ask questions on any topic.
- Study Library : Every lesson, class and PDF by programme.

© SkillNavora · skillnavora.in · Facts and tools change; check official sites before relying on a figure.

ORIGINAL CONTENT

Original AI Product Management content

The AI Product Management programme already on SkillNavora, included unchanged so the full course is in one place.

Quick links Study Library (all material) Learning sites for PM AI builder tools (how-to) Webhooks & Playwright MCP Webhook URLs and keys (start here) AI Skills Academy AI Skills Map Context engineering (PDF) AI skills & templates (PDF) Orchestration (PDF) PRD library (120 ideas) AI PRD Writer AI PM Toolkit AI Evals (PDF) Browser automation (PDF) AI second brain (PDF) PM handbook Top 100 PM questions AI product sense Q&A (86) PM practice test Practice Lab Tool tutorials AI Agents & MCP for PMs

PRD documents

PRD library: template, framing guide and 120 PRD ideas

Learn how top PMs frame a PRD, download the free template, and turn any of 120 ideas across 15 industries into a full PRD with AI.

Open the PRD library ✨ AI PRD Writer PRD template (Word) AI Evals for PMs (PDF)

Context window

Context window: what a model can see at once

The context window is the most tokens a model can consider in one request. It covers your prompt, any documents you paste, the conversation so far and the reply. Anything outside it is invisible to the model. For PMs: a longer window fits more material but costs more and can slow responses; put the most important facts near the instructions; use retrieval (RAG) rather than pasting whole libraries; and test answers on long inputs, where quality can drop.

Context engineering course Context windows in the AI PM 2 course

Tokens and context in the Grokbot course

AI-First Product Management

Resources for Product Managers

The PM Master Handbook, AI and product tools with their official websites, interview practice, a frameworks cheat sheet, a PRD template and free courses.

Open the PM handbook library: all 68 PDFs ↗ PM Glossary Handbook Tokens, Pricing and Tokenization Vibe Coding addendum

PM Master Handbook by SkillNavora
23 modules and 620+ pages: 12 Product Core modules, an 11-class AI Tools Lab, frameworks, case studies, projects and interview prep.

Open the interactive handbook ↗ Complete handbook PDF (620+ pages, 23 MB) ↗

Product Core modules

1Fundamentals of Product Management*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
2Product Strategy and Strategic Thinking*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
3Problem Discovery and Market Research*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
4Problem Framing and Prioritization*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
5UX and Design Collaboration*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
6Effective Communication and Stakeholder Management*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
7Product Analytics and Metrics*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
8Product Launch, Growth, and Monetization*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
9Tech 101 and System Design for PMs*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
10LLMs, RAGs, and AI Foundations*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
11Vibe Coding for Product Managers*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides
12Agentic AI and AI Systems*Handbook chapter, detailed PDF, handbook and class slides*PDF Detailed PDF Handbook Class Slides

AI Tools Lab

13AI Tools Landscape and Your Starter StackPDF ↗ 14Claude for Daily PM WorkPDF ↗ 15Claude Code: Basic to AdvancedPDF ↗ 16Lovable: Basic to AdvancedPDF ↗ 17Replit and Other AI App BuildersPDF ↗ 18Figma and AI Design ToolsPDF ↗ 19Notion, PRD Tools and AI DocumentationPDF ↗ 20Automation with Zapier, Make and n8nPDF ↗ 21AI Chatbots for BusinessPDF ↗ 22Voice Agents for BusinessesPDF ↗ 23Learning Paths, Courses and CommunitiesPDF ↗

Business AI Stack (chapters 24 to 31)

24The business AI stack: tools and where to learn themPDF ↗ 25Next.js: customer-facing dashboardsPDF ↗ 26Zoho CRM: managing customers and leadsPDF ↗ 27WhatsApp Business Platform: customer messagingPDF ↗ 28Deployment with Vercel and AWSPDF ↗ 29Jira: tracking product and development workPDF ↗ 30Meta Business Suite, social media marketing and Google AdsPDF ↗ 31A one-page website that generates leadsPDF ↗

Complete AI tool guides (15)

1n8n: workflows, Gmail and an AI reply agentPDF ↗ 2Make: scenarios, routers and lead routingPDF ↗ 3Zapier: invoices from Gmail to Drive and a sheetPDF ↗ 4Webhooks: how apps talk in real timePDF ↗ 5Replit: one app for the web and a phonePDF ↗ 6Lovable: a job application trackerPDF ↗ 7Figma: components, auto layout and handoffPDF ↗ 8Claude Code: an AI coding agent in your projectPDF ↗ 9MCP: connecting AI assistants to your toolsPDF ↗ 10LLMs: how they work and how to call themPDF ↗ 11RAG: answers from your own documentsPDF ↗ 12AI agents: a job-search agent with approvalsPDF

↗ 13Agentic AI: planning, guardrails and teamsPDF ↗ 14Generative AI: uses, limits and a weekly reportPDF ↗ 15Other AI tools: choosing one per jobPDF ↗

Complete AI tool guides (one PDF, 30 pages) ↗

Libraries

★Framework Library*29 frameworks with worked examples and templates*PDF ↗ ★Case Study Library*12 product case studies*PDF ↗ ★Practical Projects and Portfolio*5 end-to-end portfolio projects*PDF ↗ ★Interview Handbook*100+ PM interview questions with model answers*PDF ↗ ★AI Tools Lab*Every AI tools class with step-by-step labs*PDF ↗

PM interview practice (free, on the original sites)

Hand-picked public question banks and guides from other PM communities. Each link opens the original page in a new tab.

PM frameworks cheat sheet

Framework	Use it to	How it works
RICE	Prioritise a backlog	Reach × Impact × Confidence ÷ Effort. Score each idea, then sanity-check the ranking.
Kano	Decide which features delight	Classify features as Must-be, Performance, Attractive, Indifferent or Reverse using a short survey.
MoSCoW	Scope a release	Must have, Should have, Could have, Won't have (this time).
Jobs-to-be-done	Understand motivation	When [situation], I want to [motivation], so I can [outcome].
AARRR (Pirate metrics)	Map the growth funnel	Acquisition, Activation, Retention, Referral, Revenue.
HEART	Measure user experience	Happiness, Engagement, Adoption, Retention, Task success, each with goals, signals and metrics.
North Star metric	Align the team	One metric that captures the value users get, supported by input metrics the team can move.
Five whys	Find the root cause	Ask 'why?' repeatedly until you reach a cause you can act on.
CIRCLES	Answer product design questions	Comprehend, Identify users, Report needs, Cut through prioritisation, List solutions, Evaluate trade-offs, Summarise.

PRD template

PRD: [Feature name] Owner: [Name] · Status: Draft · Last updated: [Date] 1. Problem Who has the problem, what it is, and the evidence (data, interviews, tickets). 2. Goal and success metrics Primary metric (with baseline and target), secondary metrics, guardrails. 3. Users and use cases Target segment, top 3 user stories: "As a [user], I want [action] so that [benefit]." 4. Solution What we'll build (link designs), what's out of scope, open questions. 5. AI behaviour (if applicable) Inputs, model, prompt/approach, what it must never do, human review, evaluation set and pass criteria, cost per request. 6. Launch plan Experiment design, rollout stages, support and comms, rollback criteria. 7. Risks and dependencies Technical, legal/privacy, operational; owners and mitigations.

Free courses