AI Governance, Safety & Privacy

PM cheat sheet: turn policy into product decisions and controls

USEFUL DEFAULT FOR CONSEQUENTIAL WORKFLOWS

AI interprets → **software enforces** → **people authorize**

Use stricter controls as data, authority, or impact rises.

1 · FAST RISK SCAN

Ask these before choosing controls

- Data: What enters the model? Client-confidential, personal, regulated, credentials, or proprietary information?
- Authority: Does AI only suggest, or can it decide, send, publish, approve, change records, spend money, or touch production?
- Impact: What happens if the answer is wrong, biased, stale, leaked, or confidently unsupported?
- Access: Who may see the source data and output? Could retrieval expose information across users, teams, or clients?
- Oversight: Can a person review important outputs, inspect evidence/logs, stop the system, and recover from a bad action?

2 · PRACTICAL RISK LEVELS

Scale review to the use case

LOW	Drafting, brainstorming, public-info summaries. No sensitive data; no consequential action.	**Normal QA + approved-tool rules.**
MEDIUM	Internal knowledge, client-confidential content in approved systems, support suggestions, code assistance.	**Privacy/security review, permissions, evals, source checks, monitoring.**
HIGH	Sensitive/regulated data + material decisions, autonomous external actions, production changes, employment/financial/legal/medical impact.	**Formal approval, human authorization, least privilege, strong evals, logging, rollback/kill switch, incident plan.**

Risk is contextual. Company policy, contracts, and law can raise the required level.

3 · CONTROL MENU

Choose controls that match the risk

DATA	Minimize/redact inputs; approved enterprise tools; retention/training settings; vendor/DPA review.
ACCESS	Least privilege; tenant/project boundaries; separate read from write permissions; secrets outside prompts.
HUMAN	Human approval before consequential actions; clear escalation path; do not hide uncertainty behind confidence scores.
EVIDENCE	Ground important claims in approved sources; show provenance when users need to verify what the AI relied on.
TEST	Representative + high-risk evals; permission-boundary tests; prompt-injection/adversarial cases; safe-failure behavior.
OBSERVE	Logs/traces appropriate to the risk; monitor failures, corrections, drift, model/vendor changes, and misuse.
RECOVER	Pause/kill switch for risky actions; rollback where possible; preserve logs; incident owner + notification path.

4 · PM REVIEW FLOW

1 Intake → 2 Classify risk → 3 Data/privacy + vendor review → 4 Define authority & controls → 5 Eval/test → 6 Approve → 7 Monitor → 8 Incident/review

The PM owns the product questions; specialists own legal/security determinations.

Re-review when data, permissions, model/vendor, or impact meaningfully changes.

STOP AND ESCALATE WHEN

- A new use case introduces sensitive/regulated data or unclear data-retention/training terms.
- An agent gains new write/action permissions, especially money, production, customer messages, or record changes.
- AI output becomes a material input to a consequential decision without meaningful human review.
- You cannot explain who is accountable, what was accessed, why an action happened, or how to undo it.

Framework anchors: NIST AI RMF (Govern · Map · Measure · Manage), NIST GenAI Profile, ISO/IEC 42001.

Sep 2026

Use this as a product review aid, not a substitute for company policy, legal advice, privacy, security, or compliance review.