Seven Pillars of Trustworthy AI Implementation Guide

Type: Implementation Guide | **Audience:** AI Governance Teams, Ethics Officers

1. The Seven Pillars Overview

Pillar	Core Principle	Key Practice	Primary KPI
1. Transparency	Making AI understandable	Model documentation	% models documented
2. Fairness	Equitable outcomes	Bias testing	Disparity ratios
3. Accountability	Clear responsibility	RACI matrix	Incident resolution time
4. Safety	Preventing harm	Red teaming	Harm incidents
5. Privacy	Data protection	Data minimization	Privacy complaints
6. Human Oversight	Human control	Override mechanisms	Override rate
7. Social Wellbeing	Societal benefit	Impact assessment	Social impact score

2. Implementation Checklist by Pillar

Pillar 1: Transparency

☐ Model cards published for all AI systems
☐ AI disclosure when users interact with AI
☐ Explainability tools deployed (SHAP, LIME)

Pillar 2: Fairness

☐ Bias testing across demographic groups completed
☐ Fairness metrics defined and monitored
☐ Mitigation strategies implemented

Pillar 3: Accountability

☐ Accountable executive designated
☐ Decision rights documented (RACI)
☐ Redress mechanisms available

Pillar 4: Safety & Reliability

☐ Red team testing conducted
☐ Failure modes documented with mitigations
☐ Kill switch available

Pillar 5: Privacy

- □ Data minimization applied
- □ Privacy-enhancing technologies evaluated
- □ Consent mechanisms implemented

Pillar 6: Human Agency & Oversight

- □ Human override capability implemented
- □ Escalation paths defined
- □ Automation level appropriate to risk

Pillar 7: Environmental & Social Wellbeing

- □ Carbon footprint tracked per model
- □ Social impact assessment completed
- □ Efficiency optimization implemented

3. Maturity Assessment

Level	Description	Characteristics
Level 1: Aware	Principles documented	Ad-hoc practices, reactive
Level 2: Developing	Some practices in place	Standardized templates, manual
Level 3: Mature	Systematic implementation	Integrated processes, measured
Level 4: Optimizing	Continuous improvement	Automated, predictive, leading

Current Maturity by Pillar:

Transparency: L___ | Fairness: L___ | Accountability: L___ | Safety: L___

Privacy: L___ | Human Oversight: L___ | Social Wellbeing: L___

Organization: _________________________ **Assessment Date:** ______________

Assessed By: _________________________ **Next Review:** ______________