Seven Pillars of Trustworthy AI

Type: Comprehensive Framework

Target Audience: AI Governance Teams, Executive Leadership, Ethics Boards

Trust is the currency of AI adoption. This framework covers the seven essential pillars required for building and maintaining trustworthy AI systems. Each pillar represents a dimension that must be addressed for AI systems to be worthy of stakeholder confidence.

The Seven Pillars Overview

Pillar	Definition	Key Requirement	Failure Example
1. Safety	System does not fail dangerously	Red teaming, failure mode analysis, safe defaults	Flash Crash wiping market value
2. Fairness	Equitable outcomes across groups	Bias testing, demographic parity, disparate impact analysis	Amazon hiring tool penalizing women
3. Transparency	Users know when AI is involved	Disclosure, decision explanations, documentation	Opaque loan denials with no reason
4. Privacy	Personal data protected	Data minimization, differential privacy, consent	Clearview AI scraping without consent
5. Accountability	Clear ownership of outcomes	Designated executives, audit trails, redress	"The algorithm did it" with no owner
6. Human Oversight	Meaningful human control	Override capability, escalation paths, competent reviewers	Autonomous systems with no off switch
7. Robustness	Reliable under varied conditions	Adversarial testing, edge cases, graceful degradation	Model failing on slightly different data

Pillar Interconnections

The seven pillars are interdependent. Weakness in one undermines the others:

- **Transparency enables Accountability** - You cannot be responsible for what you cannot see
- **Fairness requires Robustness** - Systems that break easily impact vulnerable users first
- **Safety depends on Human Oversight** - Humans must be able to intervene when systems fail
- **Privacy underpins Trust** - Surveillance destroys user confidence in AI systems

Pillar Assessment Checklist

1. Safety

- ☐ Red team testing conducted
- ☐ Failure modes identified and documented
- ☐ Safe defaults implemented
- ☐ Kill switch / emergency shutdown available

2. Fairness

- ☐ Bias testing across demographic groups
- ☐ Disparate impact analysis completed
- ☐ Fairness metrics defined and monitored
- ☐ Mitigation strategies implemented

3. Transparency

- ☐ Users informed when interacting with AI
- ☐ Decision explanations available
- ☐ Model Cards / documentation published
- ☐ Audit trail maintained

4. Privacy

- ☐ Data minimization applied
- ☐ Consent mechanisms implemented
- ☐ Privacy-enhancing technologies evaluated
- ☐ Data retention policies enforced

5. Accountability

- ☐ Accountable executive designated
- ☐ Decision rights documented
- ☐ Redress mechanisms available
- ☐ Audit trail preserved

6. Human Oversight

- ☐ Override capability implemented
- ☐ Escalation paths defined
- ☐ Reviewers trained and competent
- ☐ Automation level appropriate to risk

7. Robustness

- ☐ Adversarial testing conducted
- ☐ Edge cases identified and handled

☐ Graceful degradation implemented
☐ Performance monitored in production

AI System: _____

Assessment Date: _____ **Score:** _____ / 28 pillars addressed

Assessed By: _____ **Next Review:** _____