Trustworthy AI

Fairness, Transparency, and Robustness

July 19, 2025

Outline

Introduction

Why Trustworthy AI?

As AI becomes widespread, we must ensure it is:

- Fair — avoids discrimination
- Transparent — understandable and explainable
- Robust — secure and resilient to failures

Core Pillars

Trustworthy AI is guided by 3 major principles:

1. **Fairness**
2. **Transparency (Explainability)**
3. **Robustness (Safety & Security)**

Fairness

What is Fairness in AI?

An AI system is fair if it does not produce discriminatory outcomes against individuals or groups based on:

- Race, gender, age, religion
- Sensitive or protected attributes

Fairness Metrics

Let A be a sensitive attribute and $\hat{Y}$ the predicted label.

Demographic Parity:

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1)$$

Equalized Odds:

$$P(\hat{Y} = 1|A = a, Y = y) \text{ should be equal for all } a$$

Individual Fairness:

$$\text{Similar individuals } \Rightarrow \text{ similar predictions}$$

Fairness Trade-offs

- Trade-offs exist between fairness and accuracy
- Achieving all fairness metrics at once is often impossible
- Context-specific fairness definitions must be selected

Transparency

Why Transparency?

- Stakeholders must understand how AI makes decisions
- Lack of transparency reduces trust
- Enables accountability and debugging

Types of Explainability

- **Global explanations**: model-level understanding
- **Local explanations**: explain individual predictions
- **Post-hoc**: use external tools (e.g., LIME, SHAP)
- **Interpretable models**: inherently simple (e.g., decision trees)

Interpretable vs Black-box Models

Model Type	Interpretability	Examples
Linear Models	High	Logistic Regression
Decision Trees	High	CART, C4.5
Neural Nets	Low	CNNs, RNNs
Ensembles	Low	Random Forest, XGBoost

Robustness

What is Robustness?

Robustness refers to an AI system's ability to remain:

- Accurate under noisy, adversarial, or out-of-distribution inputs
- Stable when small changes are made to input
- Resilient against manipulation and attacks

Types of Robustness

- **Adversarial robustness** — resisting malicious perturbations
- **Distributional shift** — generalizing to new data domains
- **Resilience to missing/noisy data**

Simple Adversarial Example

Let x be an image input and $\hat{y} = f(x)$. We add perturbation δ, where $\|\delta\|$ is small:

$$\hat{y} = f(x), \quad \hat{y}' = f(x + \delta), \quad \hat{y} \neq \hat{y}'$$

- Model prediction flips even with imperceptible change
- Needs robust training to defend

Best Practices

Best Practices for Trustworthy AI

- Use interpretable models when possible
- Audit datasets for bias
- Evaluate fairness using metrics
- Perform robustness testing (e.g., adversarial attacks)
- Document model behavior clearly

Key Takeaways

- Trustworthy AI ensures fairness, transparency, robustness
- Must balance ethical principles with performance
- Requires tools, metrics, and careful design

Thank you!

Questions?