

# Understanding Trustworthy AI

## A Beginner's Guide to the 7 Key Characteristics

As AI becomes woven into our daily lives—recommending movies, assisting medical diagnoses, and more—we need to trust these complex systems. But what makes AI worthy of our trust? Discover the seven essential qualities that ensure AI is beneficial, safe, and fair.

Swipe to explore each characteristic →



# The Foundation of Trust

NIST identifies seven interconnected characteristics that build trustworthy AI.

Two play special roles:

## Valid and Reliable

The necessary foundation—if AI doesn't work correctly, nothing else matters

## Accountable and Transparent

The overarching pillar that supports and verifies all other characteristics

These characteristics aren't separate checkboxes—they're deeply interconnected, often requiring careful balancing. Let's explore each one.

# Valid and Reliable

Does It Work Correctly  
and Consistently?

## Validity

The AI fulfills its intended purpose  
and gives you the correct result

## Reliability

The AI performs as required without  
failure over time, consistently

**Think of a GPS app:** It's **valid** because it gives correct routes, and **reliable** because it does so every time—city or countryside. This foundation includes accuracy (correct predictions) and robustness (performance in varied conditions).





# Safe

## Will It Cause Harm?

Safety means an AI system should not, under defined conditions, endanger human life, health, property, or the environment. It's about actively preventing harm, whether intentional or unintentional.

01

---

### Design for Safety First

Employ safety considerations from the beginning of planning, not as an afterthought

02

---

### Enable Human Intervention

Ensure humans can intervene or shut down a system that's not behaving as expected

Like car safety features such as automatic braking, safe AI monitors and intervenes to prevent accidents.

# Secure and Resilient

Can It Withstand Attacks and Failures?

## Security

Protection from threats and preventing unauthorized access

- Data poisoning attacks
- Adversarial examples
- Unauthorized use prevention

## Resilience

Ability to handle and recover from adverse events

- Graceful degradation
- Backup systems
- Recovery mechanisms

**Think of a bank vault:** Thick walls and alarms provide *security*, while backup locks and secondary systems ensure *resilience* if one measure fails.

# Accountable and Transparent

Can We See How It Works and Who Is Responsible?



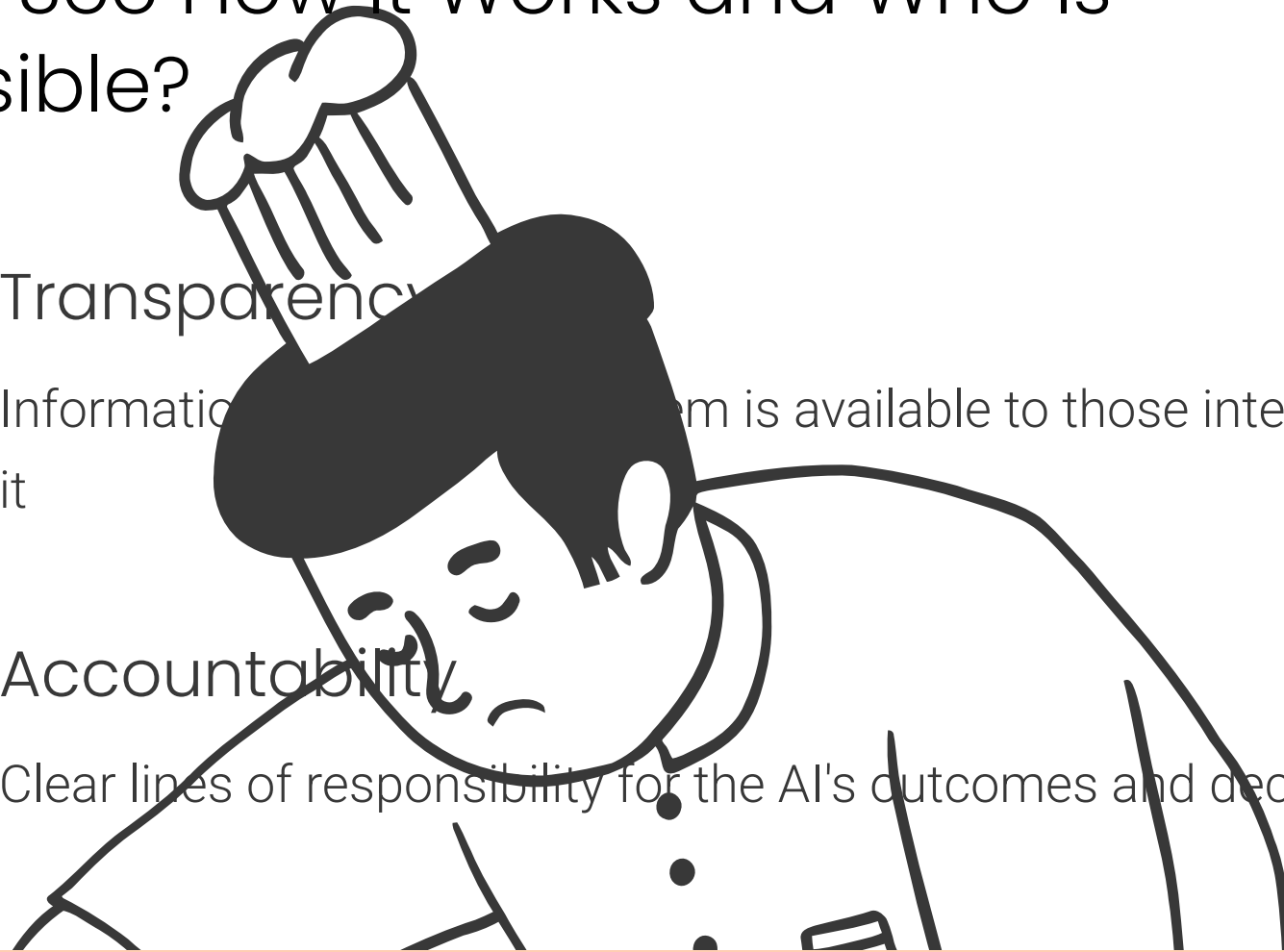
Transparency


Information about how the system works is available to those interacting with it





Accountability

Clear lines of responsibility for the AI's outcomes and decisions



 **Key Insight:** Accountability cannot exist without transparency—you can't hold someone responsible if you don't know how a decision was made. Meaningful transparency provides the right level of information to the right person.

If an AI denies you a  about the  factors used, the only way to seek **actionable redress** and hold organizations accountable for potentially unfair decisions.

# Explainable and Interpretable

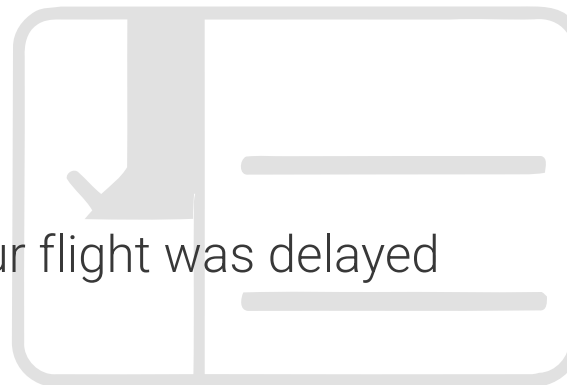
## Can We Understand Its Decisions?

If transparency answers "what happened," these characteristics answer "how" and "why." Here's the crucial distinction:

1 Transparency

### **What happened?**

A flight status board shows your flight was delayed



2 Explainability

### **How did it decide that?**

The system shows a mechanical issue was detected during pre-flight checks

3 Interpretability

### **Why does it matter to me?**

An agent explains you'll miss your connection and are being rebooked on the next flight

These qualities help people make sense of and appropriately contextualize an AI system's output.

# Privacy-Enhanced

## Does It Protect Personal Information?

Privacy in AI refers to norms and practices that safeguard human autonomy, identity, and dignity. This involves protecting personal information and ensuring individuals control how their data is used.

### New Privacy Risks

AI can infer previously private information from seemingly non-sensitive data

### Privacy-Enhancing Technologies

Special tools (PETs) help build systems that respect privacy from the design phase

Like HIPAA rules in a doctor's office, AI privacy protections ensure sensitive information stays confidential and is only used for intended purposes.

# Fair

## Does It Manage Harmful Bias?

Fairness involves a commitment to equality and equity by actively addressing harmful bias and discrimination. Because AI learns from data, it can reflect—and amplify—biases in our society.

"While bias is not always negative, AI systems can potentially increase the speed and scale of biases and perpetuate and amplify harms."



### Systemic Bias

Pre-existing societal biases reflected in training data—the AI inherits, not creates, these biases



### Computational Bias

Biases from model errors or unrepresentative data samples



### Human-Cognitive Bias

Biases in how people perceive, interpret, or use AI information

# The Balancing Act

## Navigating Tradeoffs in Trustworthy AI

These seven characteristics are interconnected. Sometimes, improving one makes it harder to achieve another, creating a need for careful balancing based on context.

### Movie Recommendations

**Low stakes:** Prioritize privacy over detailed explanations

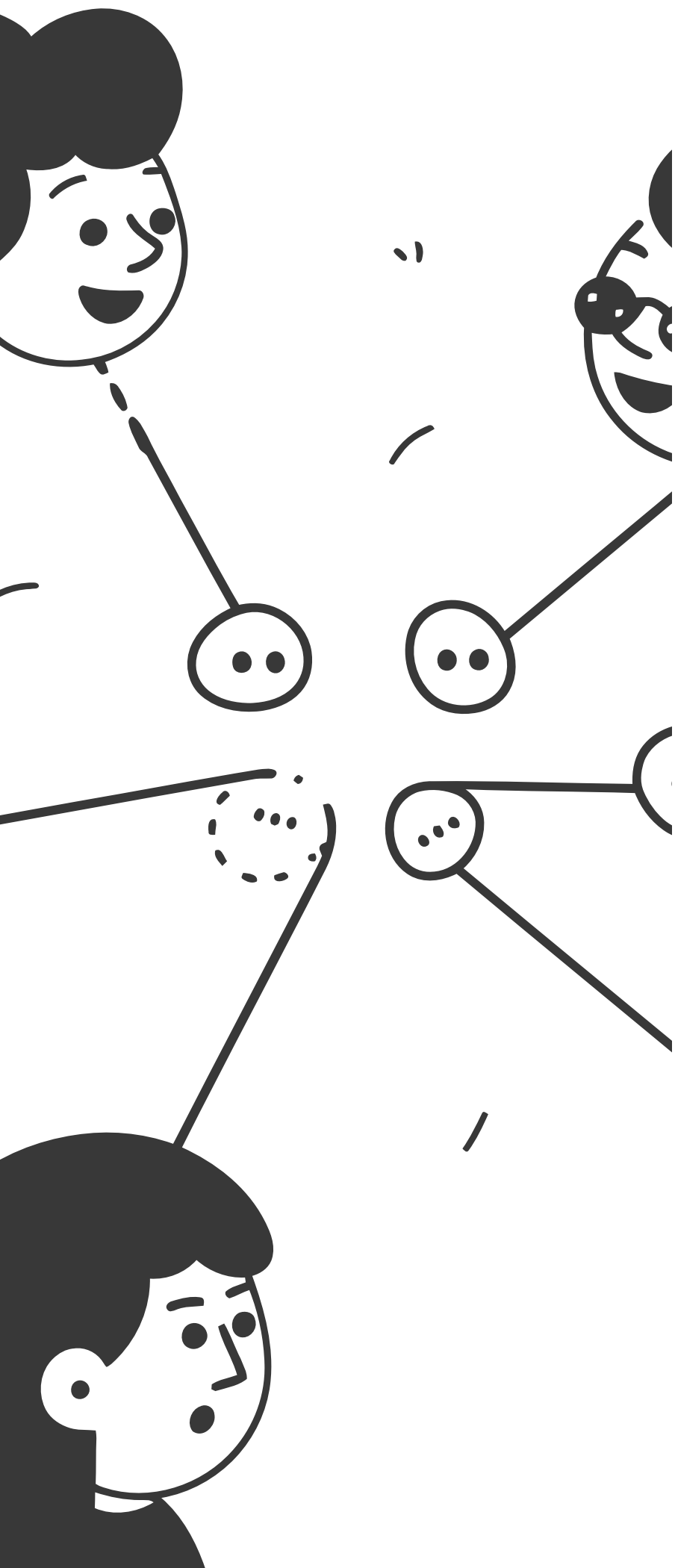
### Medical Diagnoses

**High stakes:** Prioritize accuracy and safety above all, even if less explainable

### Criminal Justice

**Critical stakes:** Prioritize fairness and explainability to ensure due process


These decisions aren't purely technical—they're ethical and societal. It's the joint responsibility of everyone in the AI lifecycle to navigate these tradeoffs transparently and justifiably.



## Why These Characteristics Matter

Understanding these seven characteristics is the foundation for responsible citizenship in an AI-driven world. Whether you're a creator, user, or someone impacted by AI, you now have the language to ask critical questions:

- Is it working correctly?
- Is it safe?
- Is it fair?
- Can we understand it?
- Does it protect privacy?

 **Building trustworthy AI is not someone else's job—it's a continuous, collective effort.** Your understanding is the first and most vital step in shaping a future where AI aligns with our most important human values.

Share this guide with someone who needs to understand trustworthy AI