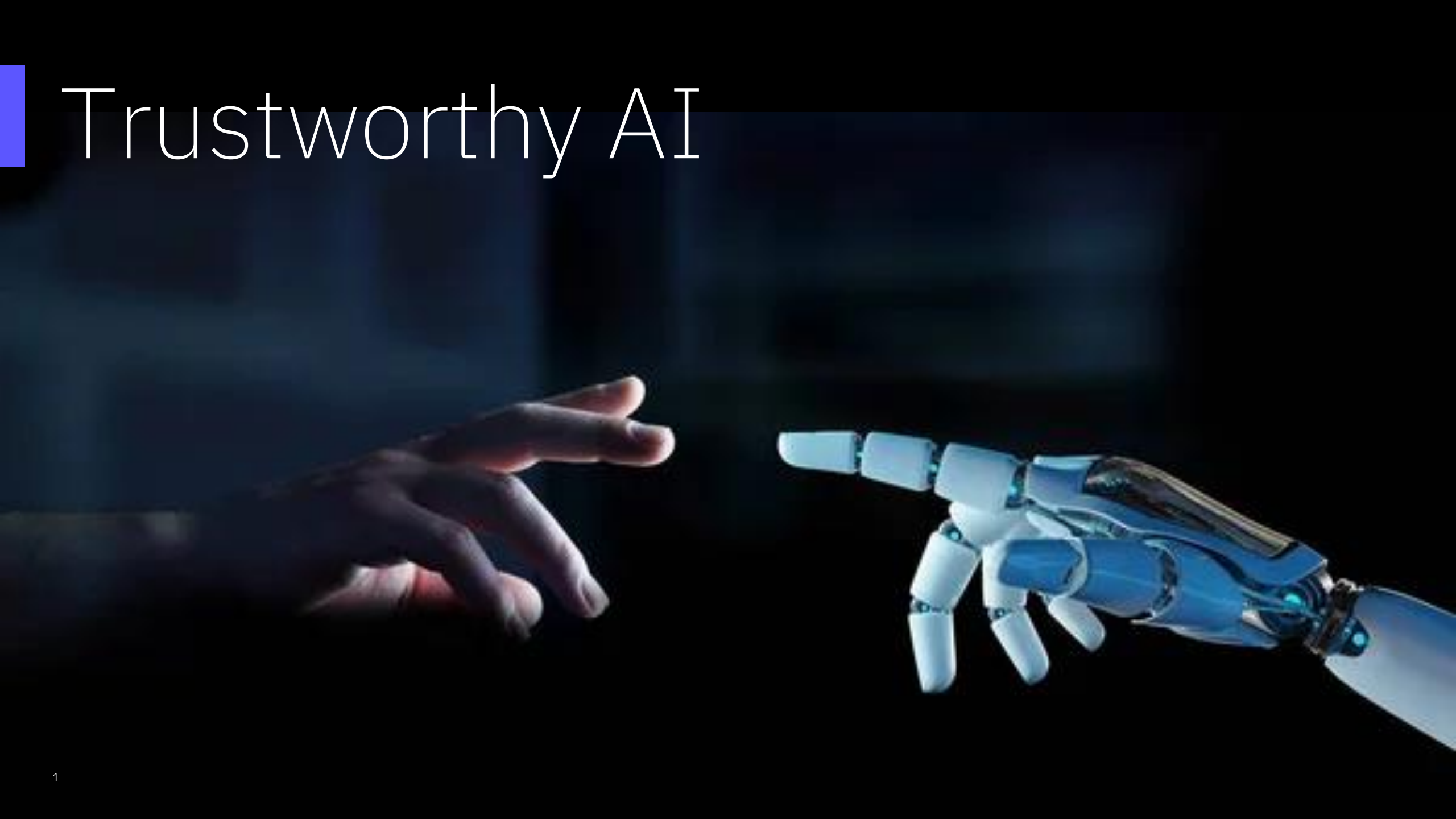
Trustworthy AI



1. The Five Pillars
2. Fairness
3. Robustness
4. Privacy
5. Explainability
6. Resources

The 5 Pillars of Trustworthy AI

Trustworthy AI

According to Morning Consult, 77% of global IT professionals report that it is critical to their business that they can trust the AI's output is fair, safe and reliable.

5 Pillars


Fairness


Robustness


Privacy


Explainability


Transparency

What is Fairness in AI?

Fairness in AI ensures that the system's outputs do not disproportionately benefit or harm specific individuals or groups. This pillar demands equitable treatment across demographic boundaries such as race, gender, age, and socioeconomic status. When fairness is neglected, AI systems can reinforce historical biases embedded in their training data.

- AI outcomes should not favor or penalize based on sensitive attributes.
- Detecting and correcting bias is critical during training and evaluation.
- Fairness can involve statistical parity, equal opportunity, and demographic parity.
- Tools: IBM AI Fairness 360, Google’s What-If Tool.

Real-world Fairness Failures and Fixes

Amazon's AI Hiring Tool (2018)

The algorithm penalized resumes with the word “women’s,” showing bias against female candidates.

Reference: Reuters, 2018

COMPAS Recidivism Risk Tool

Used in U.S. courts; found to be twice as likely to label Black defendants as high risk falsely.

Reference: ProPublica, 2016

What is Robustness in AI?

Robust AI systems remain stable and functional when exposed to noisy data, unexpected inputs, or adversarial conditions. A robust model doesn't break down under pressure — whether due to missing data, changing environments, or attacks. It’s essential for systems deployed in safety-critical applications like autonomous driving or medical diagnostics.

Resilience to noise, drift, adversarial examples.

Evaluated through stress testing, input perturbations.

Tools: CleverHans, Adversarial Robustness Toolbox (ART).

Common metric: performance under input corruption.

Real-world Robustness Challenges

Tesla Autopilot Crashes

Tesla’s AI-driven Autopilot failed in specific lighting and traffic conditions, leading to fatal accidents.

Reference: PBS News, 2024

Adversarial Panda Image (Goodfellow et al.)

Adding imperceptible noise changed an image classification from “panda” to “gibbon.”

Reference: Explaining and Harnessing Adversarial Examples, 2014

What is Privacy in AI

AI models often rely on sensitive user data. Ensuring privacy means safeguarding that data from unauthorized access or inference attacks.

This includes adopting secure data handling, minimizing retention, and enabling users to control how their data is used — all while complying with data protection regulations.

- Enforce anonymization, differential privacy, and encryption.
- Federated learning allows local training without data centralization.
- Align with GDPR, HIPAA, CCPA.
- Tools: TensorFlow Privacy, OpenMined.

Real-world Privacy Violation

Strava Fitness App Heatmap Leak (2018)

- Published global heatmaps from fitness trackers revealed locations of secret U.S. military bases.
- Reference: The Guardian, 2018

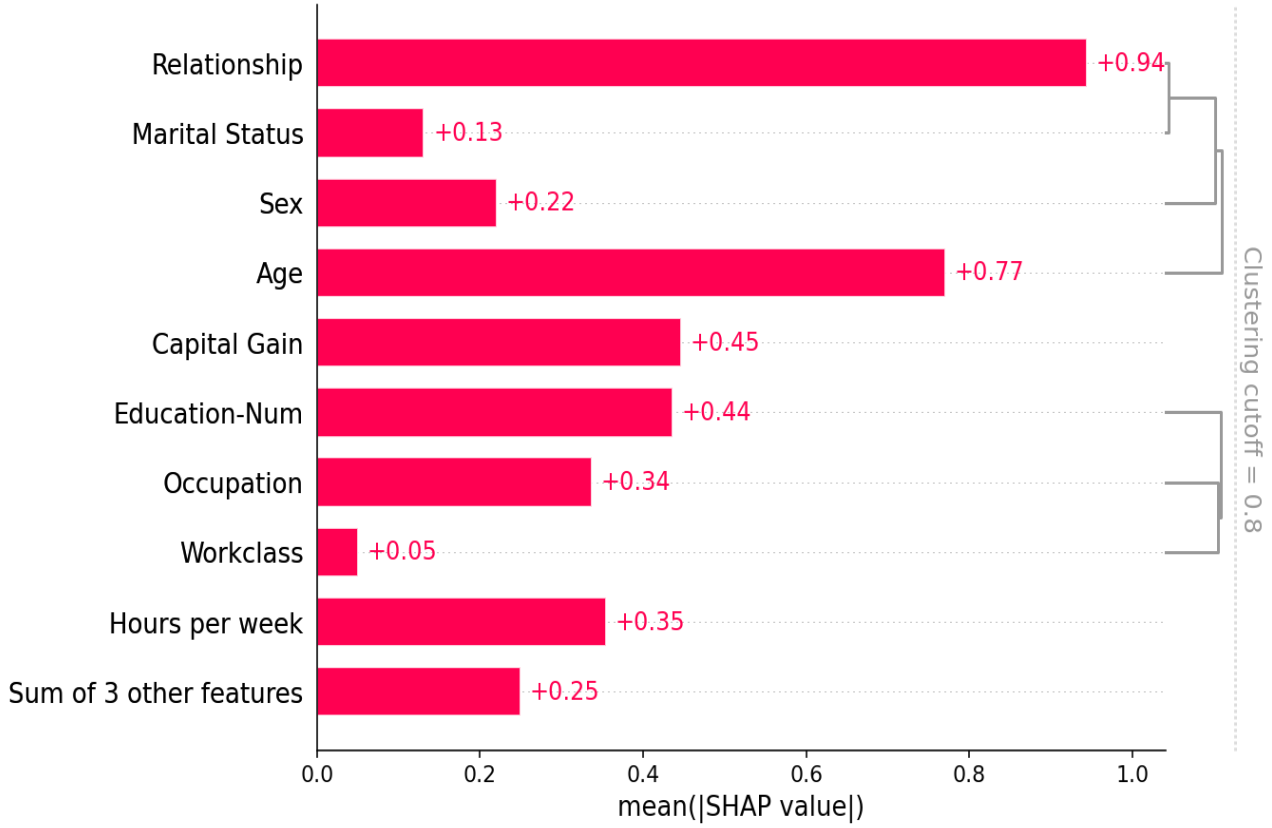
Zoom's AI Companion and Data Consent Backlash (2023)

- Zoom's update stated it could use customer data to train AI, causing public uproar and quick retraction.
- Reference: TechCrunch, 2025

What is Explainability in AI?

Explainability refers to the degree to which humans can understand and trust the logic behind an AI's decisions. Without it, users are left in the dark about why an outcome was produced, which hinders accountability and debugging. Especially in high-stakes domains like finance or healthcare, explainability builds user confidence and regulatory compliance.

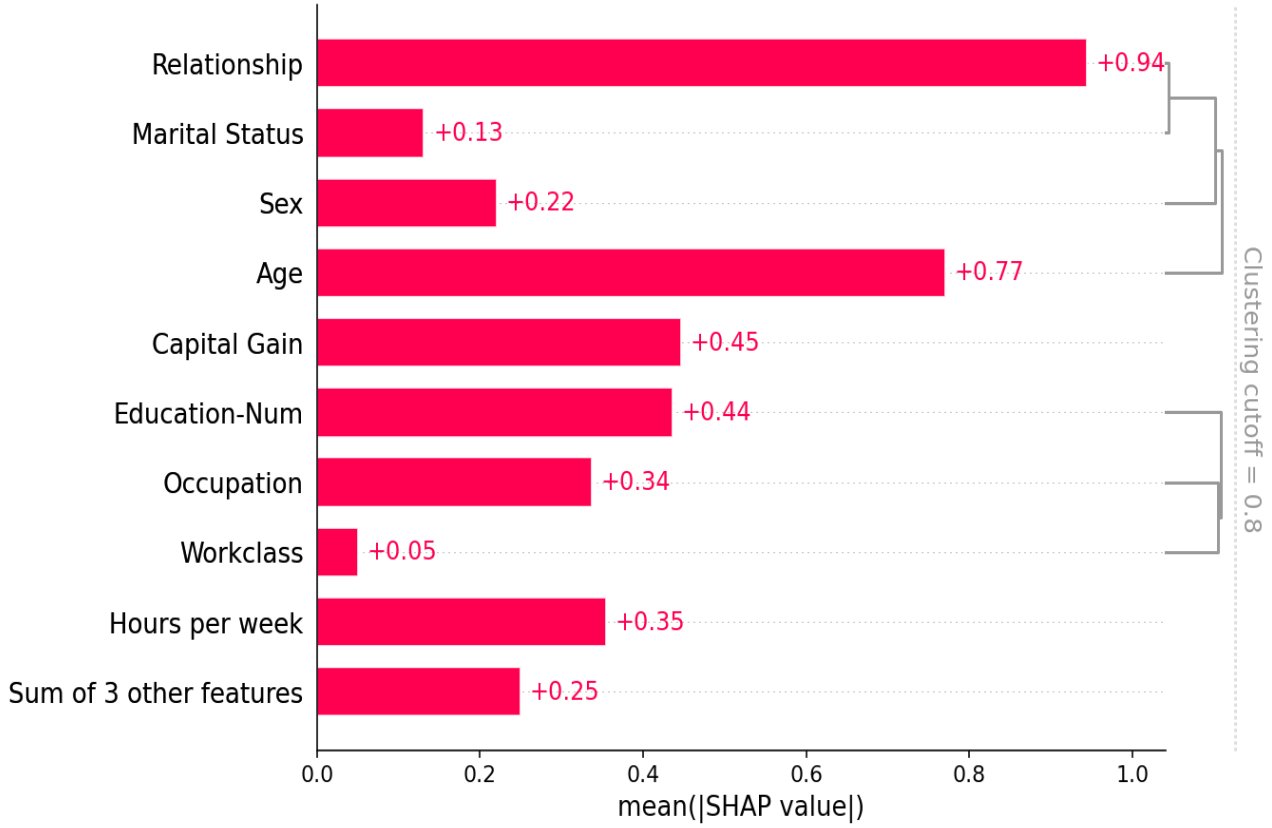
- Helps users trust and contest decisions.
- Black-box models (deep learning) need interpretability tools.
- Techniques: SHAP, LIME, counterfactual explanations.
- Useful for model debugging and audit trails.



Explainability Case Studies

Apple Card Gender Bias (2019)
Customers (e.g., Steve Wozniak and wife) got vastly different credit limits with no explanation.
Reference: CNN business, 2019

Watson for Oncology Failures
IBM Watson recommended unsafe cancer treatments, but clinicians couldn't interpret why.
Reference: Advisory Board, 2017



What is Transparency in AI

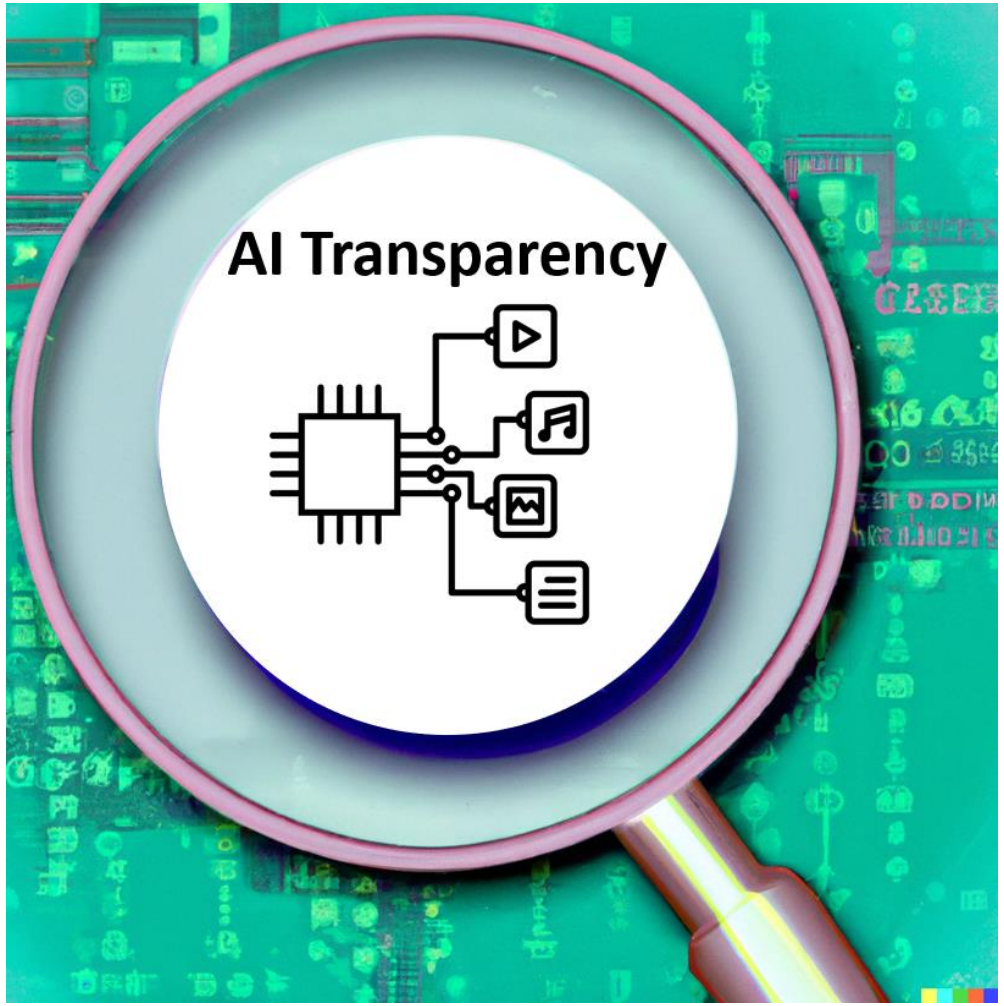
Transparency in AI means making the development, intentions, data sources, and limitations of AI systems visible and understandable to stakeholders. This openness fosters accountability and allows others — researchers, regulators, or end-users — to assess and challenge the system's design and behavior.

Document model training, assumptions, and limitations.

Disclose datasets, APIs, and decision logic when possible.

Use model cards and datasheets for datasets.

Transparency ≠ full open-source, but rather informed access.



Real Transparency Issues

Facebook's News Feed Algorithm Controversy

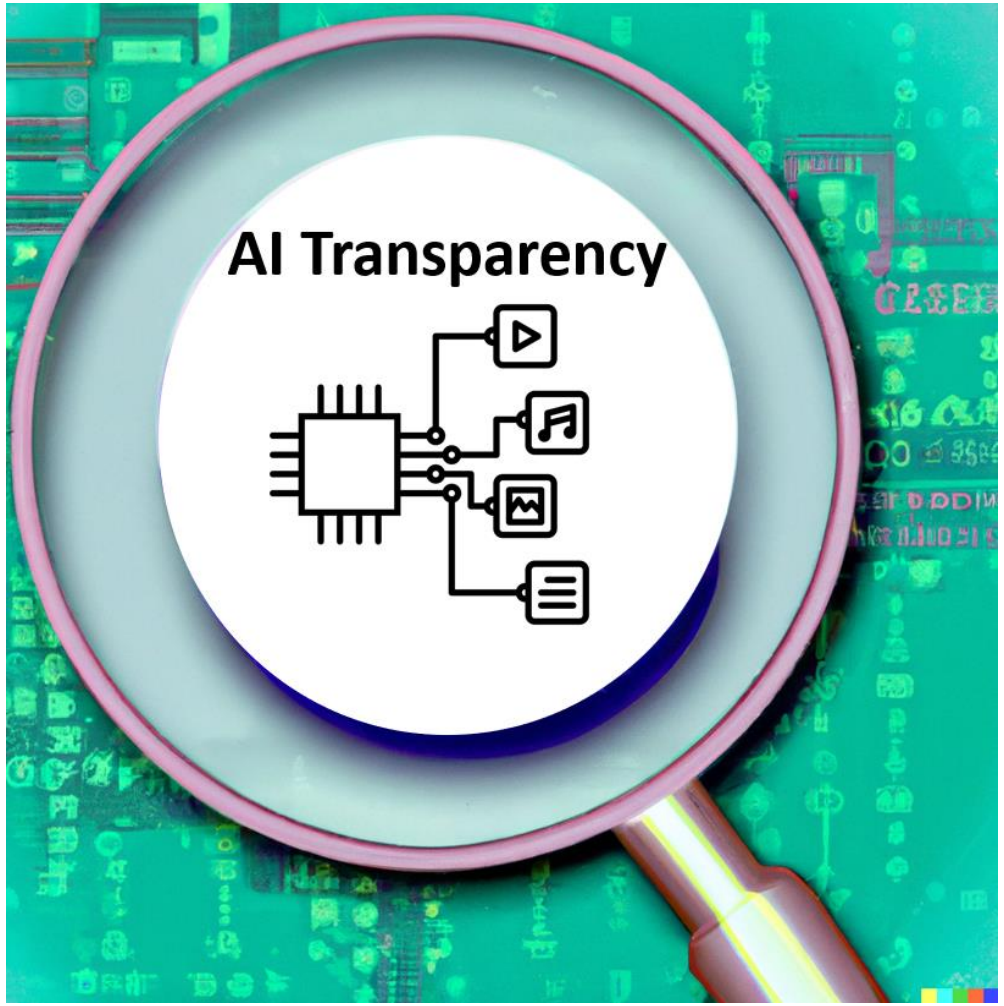
Lack of clarity on how Facebook’s algorithm amplifies certain content led to misinformation concerns.

Reference: Global Witness, 2022

Google AI Ethics Team Dismissals (2020–2021)

Firing of Timnit Gebru and Margaret Mitchell after raising concerns about AI transparency triggered global debates on accountability.

Reference: The Washington Post, 2020



Discussion topics for Week 8

Fairness:
Can an algorithm ever be “completely” fair? How do we decide what fair means in a global context?
Would removing all demographic information (e.g., race, gender) from a dataset make a model fair? Why or why not?

Robustness:
Would you feel comfortable trusting AI to fly a plane or diagnose a disease? What kinds of robustness testing would be essential?
Is it ethical to release AI systems to the public if they aren’t provably robust under real-world conditions?

Privacy
Would you consent to your health data being used to improve a medical AI model? What conditions would you want in place?
Can AI privacy protections and data utility coexist, or is it always a trade-off?

Explainability
If you can’t explain an AI decision, should you use that model in critical situations?
Can complex models (like deep learning) ever be fully explainable? Should we prioritize simpler models if they are more interpretable?

Transparency
When might full transparency conflict with trade secrets or national security? How do we balance that?
Would a "nutrition label" for every AI model help improve trust? What should it include?

Resources

Tesla Auto-pilot crashes

Facebook’s Newsfeed Algorithm Controversy