Responsible AI & ML Fairness at Google

Dr. Christoph Mittendorf | 01 September 2022

Agenda

- **01.** Ethics overview - Reliability & Fairness
- **02.** AI Principles & Responsible AI
- **03.** Transparency with Tools & Education
- **04.** Key Learnings

Google Cloud

AI Systems

Safety & Ethics overview

“AI systems can only benefit the world if we make them reliable and fair.”

Source: Google - https://blog.google/technology/ai/update-work-ai-responsible-innovation/

“AI systems can only benefit the world if we make them reliable and fair.”

Reliability is the overall consistency of a measure - it produces similar results under consistent conditions.

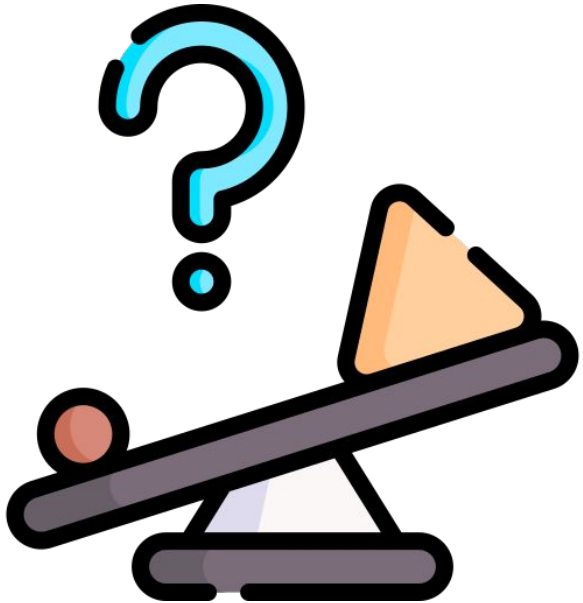
Fairness refers to the attempt of correcting bias.

*As it is the case with many ethical concepts, definitions of fairness and bias are always controversial.

In order to reach Fairness - we have to get rid of Bias!

Types of Bias

- automation bias
- confirmation bias
- coverage bias
- experimenter's bias
- in-group bias
- group attribution bias
- **human bias**
- implicit bias
- **non-response bias**
- out-group homogeneity bias
- **reporting bias**
- **sampling bias**
- **selection bias**



Example: Human bias | Reporting bias

People have different perceptions | ML is not objective

>>> Machine learning models are not inherently objective.

>>> Machine learning models are stochastic and work with probability.

>>> Improving Fairness in Classifiers is a necessity.

>>> Take proactive steps to mitigate Bias.

You are the worst example of a puppy I've ever seen.

0.07

What a sweet puppy, I want to hug her forever!

0.93

Key takeaway

Fairness is not static

"While we still have a lot to learn—and will continue learning given the dynamic and evolving nature of technology and society—we remain committed to sharing our progress and findings."

Source: Google - 2021 AI Principles Progress Update

Key takeaway

Reliability may help us to identify and combat bias.

"Designing systems to address these biases is challenging, and requires careful consideration not just of the technology, but of the societal context in which it will be deployed. But well-designed, thoroughly vetted AI systems can limit unfair bias."

Source: Google - 2021 AI Principles Progress Update

AI Principles

An ethical charter to guide the development and use of AI

“A reliable and fair AI System, like all technology, needs to be built and used responsibly.”

Source: Google - https://blog.google/technology/ai/update-work-ai-responsible-innovation/

Google AI Principles

AI should:

1. be **socially beneficial**
2. avoid creating or reinforcing **unfair bias**
3. be built and tested for **safety**
4. be **accountable** to people
5. incorporate **privacy** design principles
6. uphold high standards of **scientific excellence**
7. be made available for uses that accord with these principles

Applications we will not pursue:

1. likely to cause overall **harm**
2. principal purpose to direct **injury**
3. **surveillance** violating internationally accepted norms
4. purpose contravenes **international law** and **human rights**

2018-today

“Google has a central team dedicated to ethical reviews of AI research and new applications before launch in alignment with our principles”

Source: Google - https://blog.google/technology/ai/update-work-ai-responsible-innovation/

Responsible Innovation team - Review Process*

01 

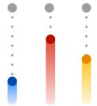
Intake

Any team can request AI Principles advice. Reviewers identify relevant AI Principles as frameworks for action.

02 

Analysis

Reviewers analyze the scale and scope of a technology's potential benefits and harms. Reviewers consult with internal experts on privacy, security, ML fairness, and other domains as needed.

03 

Adjustment

Reviewers recommend technical evaluations (e.g., checking for unfair bias in ML models).

04 

Decision

Reviewers decide whether to pursue or not pursue the AI application under review.

*Each review is unique. This is intended only as a very high-level summary, and reflects the current process.

Source: Google - https://ai.google/responsibilities/review-process/

Key takeaway

“Principles that remain on paper are meaningless.”

Putting our principles into practice is key.

-Sundar Pichai

Source: Google - 2021 AI Principles Progress Update

Transparency
with Tools & Education

Tools and education

We're building tools and resources to provide model transparency in a structured, accessible way.

Source: Google - https://cloud.google.com/responsible-ai

Three examples: Tools and education

01

Explainable AI

Explainable AI is a set of tools and frameworks to help you understand and interpret predictions made by your machine learning models to help detect and resolve bias, drift, and other gaps in data and models.

02

Model Cards

Model cards may have the potential to help investigate We recommend that released models be accompanied by documentation detailing their performance characteristics to encourage a transparent model reporting.

03

TensorFlow open-source toolkit

Responsible AI practices can be incorporated at every step of the ML workflow from (1) Problem statement (2) Data preparation, (3) Model training, (4) Model evaluation, and (5) Model deployment and monitoring.

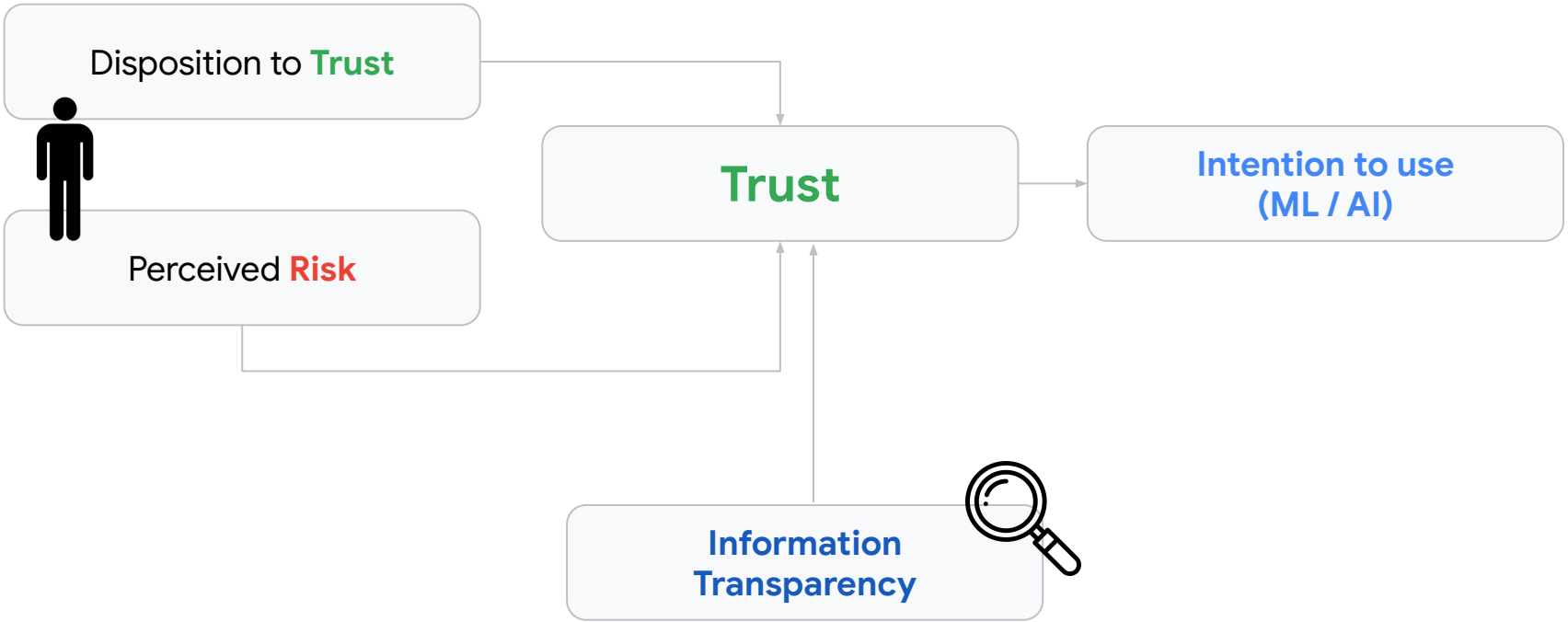
Source: Google - https://modelcards.withgoogle.com/about | https://cloud.google.com/explainable-ai | https://www.tensorflow.org/responsible_ai

The Importance of Trust

Trusting AI systems is cited as the biggest barrier for enterprises moving from the lab to production

Sources: CCS Insights IT Decision-Maker Workplace Technology Survey 2019 (pdf)

Trust and its antecedents



Source: Mittendorf et al. The Impact of Information Transparency on Trust, (2021).

Key takeaway

Information Transparency increases Trust.

Responsible AI tools are an increasingly effective way to inspect and understand AI models which leads to enhanced trustworthiness of AI Systems.

Source: Google - https://cloud.google.com/responsible-ai

Key Learnings

Four Key learnings

Fairness

Fairness is not static.

Reliability

Reliability may help us to identify and combat bias.

AI Principles

Principles that remain on paper are meaningless.

Trust

Information Transparency increases Trust.