Unpacking the blackbox: Transparency

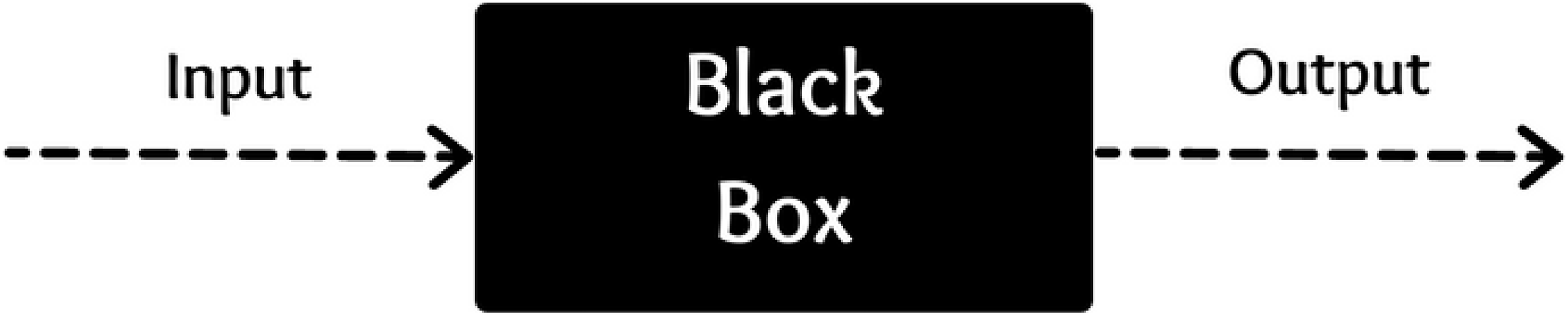
AI ETHICS

Joe Franklin
Llama Enthusiast

datacamp

Black-box nature

- AI implementations are often black boxes
- A black box in AI:
  - Known **inputs** and **outputs**

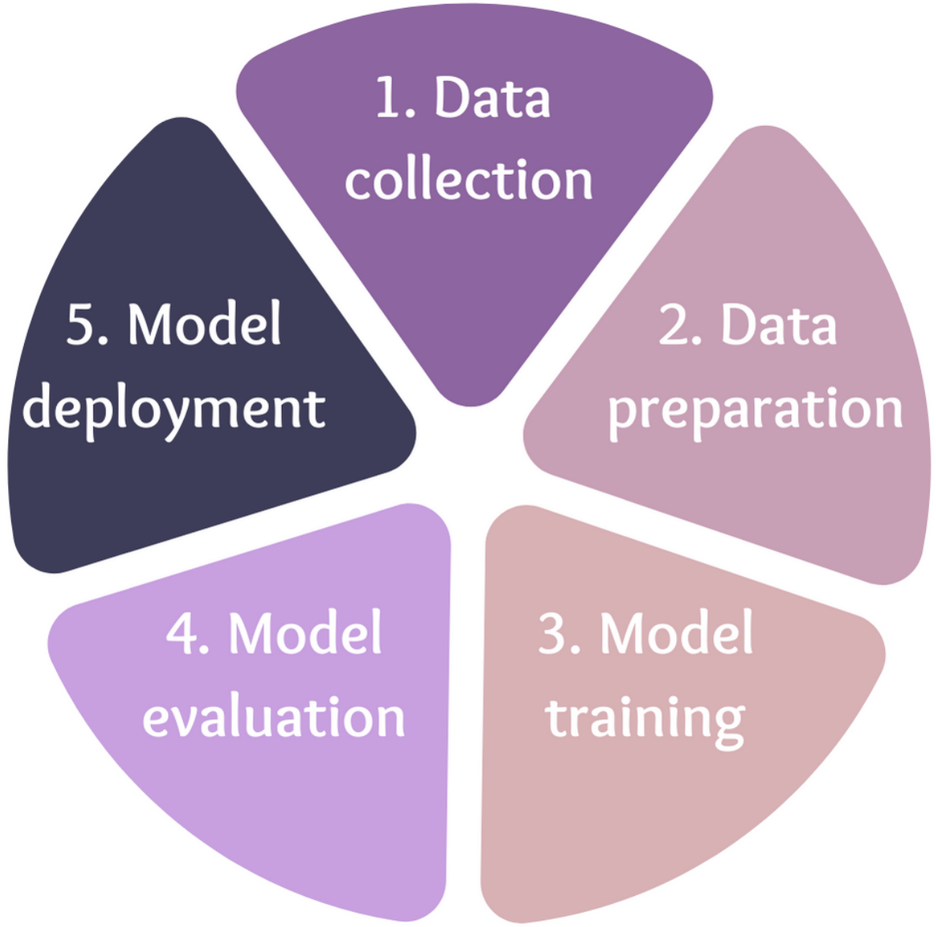


Ambiguousness is non-ideal

- Ambiguity in AI: Ethical challenge
- Question of trust:
  - Can we validate AI decisions without understanding them?
- Transparency:
  - Making an AI's decision-making process understandable
- Example:
  - Factors in AI sales model

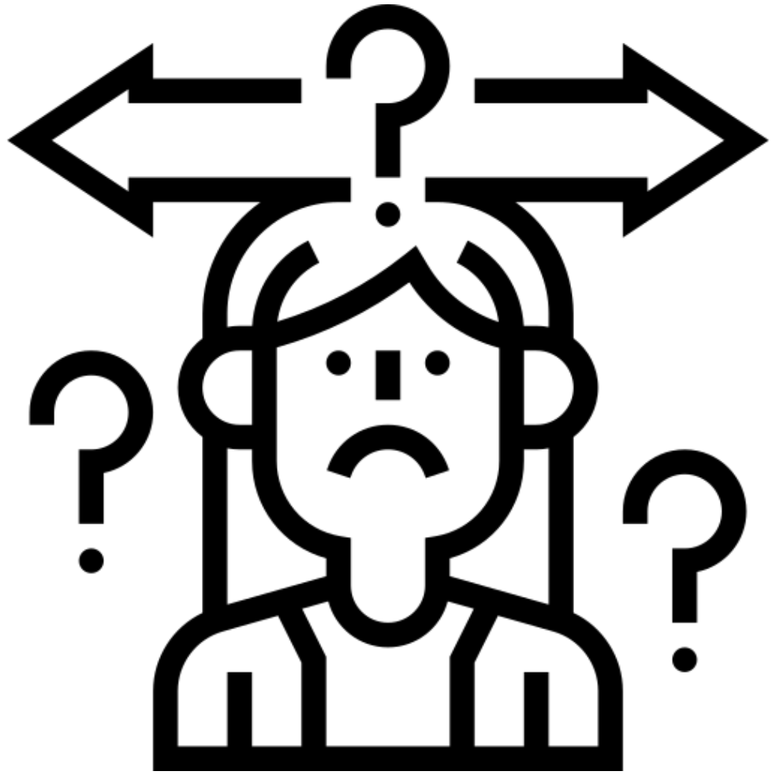
Throughout the AI life cycle

- Transparency in AI involves all stages of the AI life-cycle
- Purpose:
  - Understand the workings of the AI system
  - Gauge comfort level with its operation



A deciding factor

- Current state:
  - Transparency in AI is uncommon
  - **Hesitation** in AI adoption
- Future implications:
  - Transparency will become a **deciding factor** in users' choice of AI systems
- Actionable:
  - Organizations should prioritize transparency



[1] Icon made by Eucalyp from www.flaticon.com

Openness is key

- Openness about AI challenges and learnings is **key**
- **Transparency** encourages **innovation** in AI
- It leads to more advanced, reliable AI systems

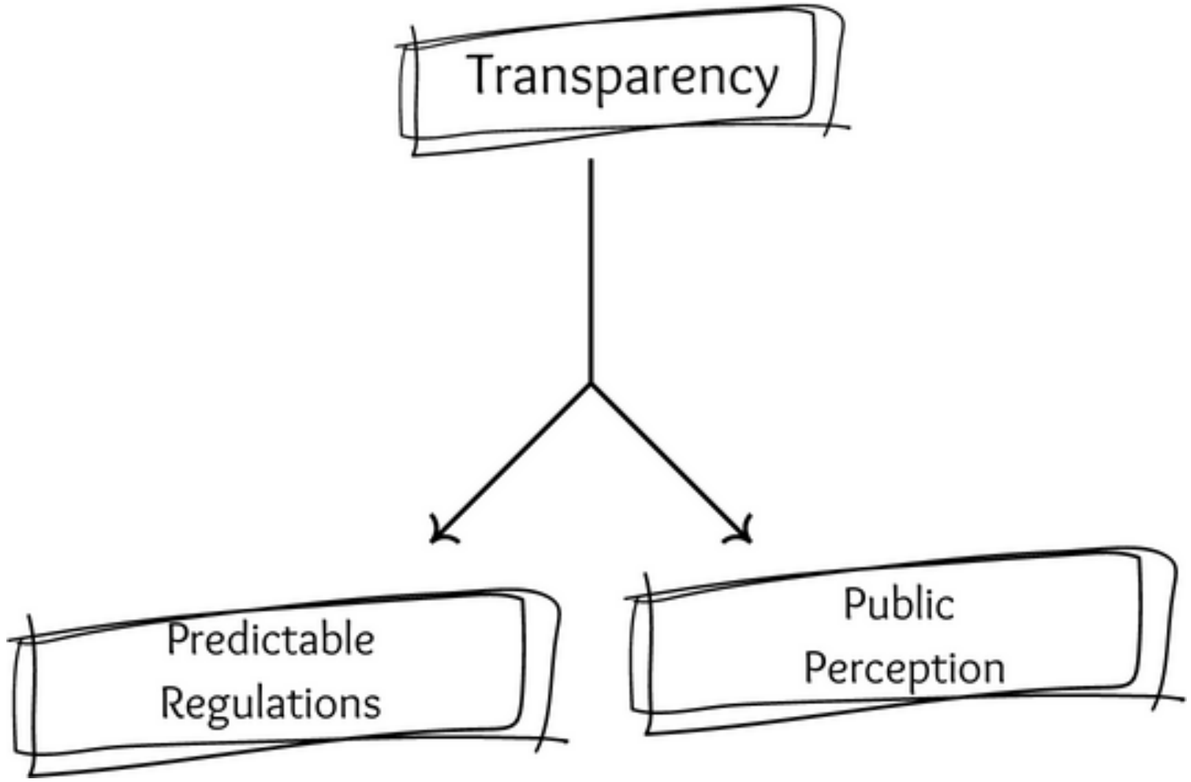
[1] Icon made by Freepik from www.flaticon.com

Embracing transparency in AI

- Transparency in AI can be intimidating but is beneficial for businesses

- Transparency leads to **predictable regulations** and **public perception**

- Companies can compete based on **strengths, culture, customer relationships** rather than secrecy



Let's practice!

AI ETHICS

AI fairness: not just a dream

AI ETHICS

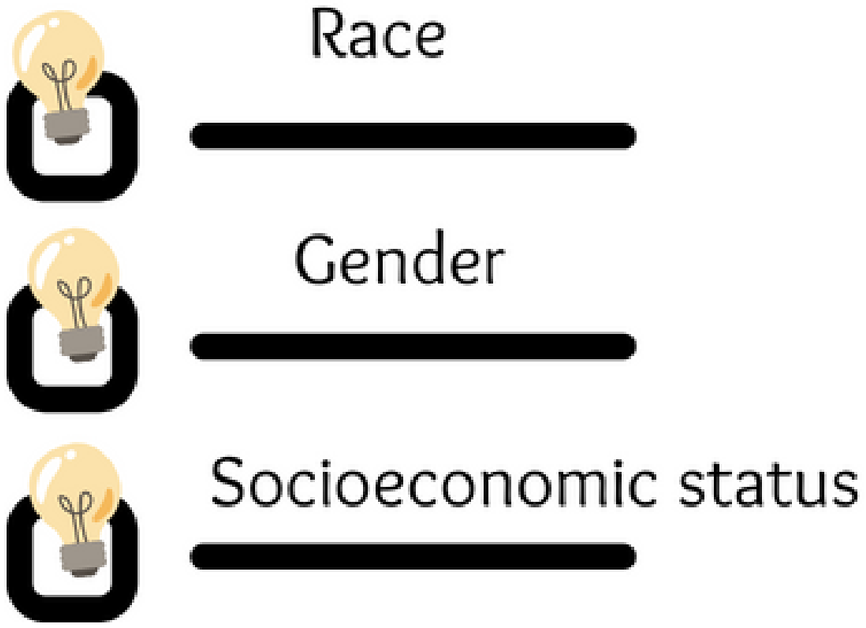
Joe Franklin

Associate Data Literacy and Essentials Manager, DataCamp

datacamp

Fairness in AI

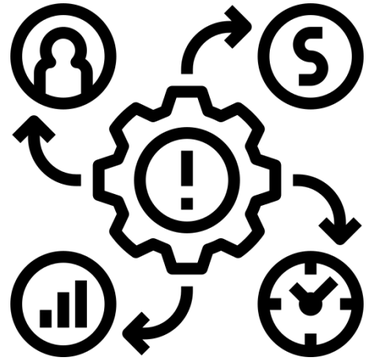
- Fairness: Ensure no group is favored over another
- Concerns race, gender, socioeconomic status, etc.

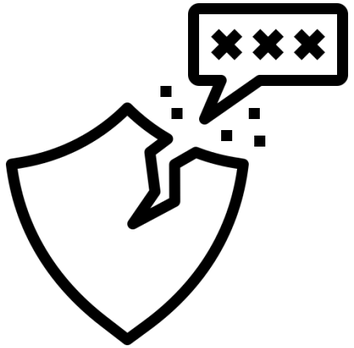


- AI should predict patient outcomes equitably
- There should be no bias towards any specific group

Why does fairness matter?

- AI's rapid processing can result in large-scale impacts





- Fairness prevents negative targeting of vulnerable populations
- Essential for responsible AI implementation, ensures equitable consideration for all

[1] Icons made by noomtah & Parzival' 1997 from www.flaticon.com

Promoting fairness

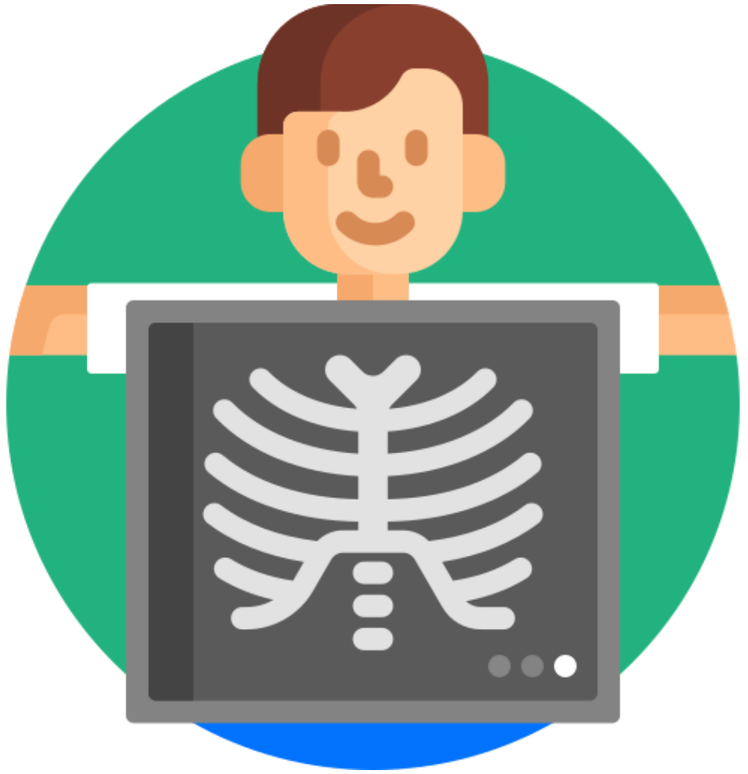
- Fairness promotion is challenging but possible
- Reduces potential bias by omitting certain variables
- Variables include race, gender, age, socioeconomic status, sexual orientation, religion

Fairness through unawareness

Unintentional issues exist

- Even with unawareness, unintentional bias can still occur

- Robust strategies needed to ensure fairness



[1] Icons made by Freepik from www.flaticon.com

Minimizing bias

- The main objective of AI fairness is minimizing **bias**
- The first step is **acknowledging** bias exists
- Remain **skeptical** and **vigilant** of AI
- Conduct frequent monitoring and audits for fairness

Let's practice!

AI ETHICS

Safeguarding AI: Accountability

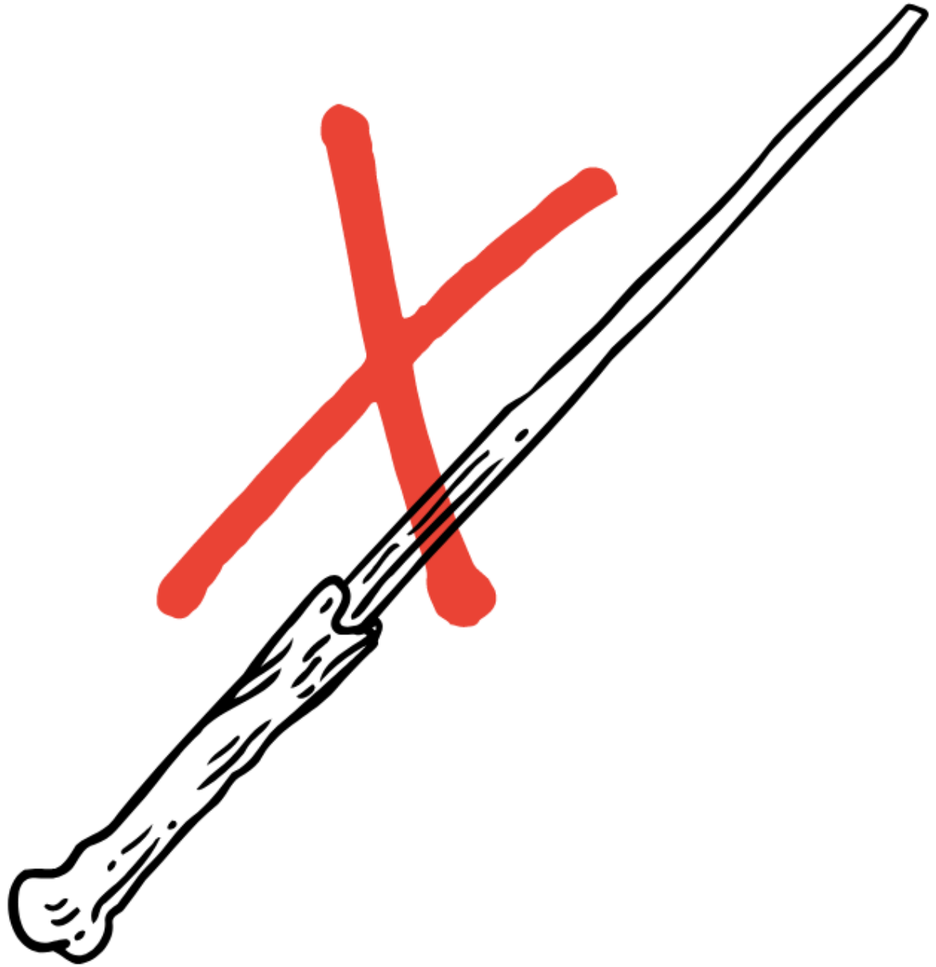
AI ETHICS

Joe Franklin

Associate Data Literacy and Essentials Manager, DataCamp

Define accountability

- Accountability:
  - Assigning responsibility for AI outcomes
- Critical in AI's development, deployment, and use
- AI isn't a responsibility-evading "magic wand"

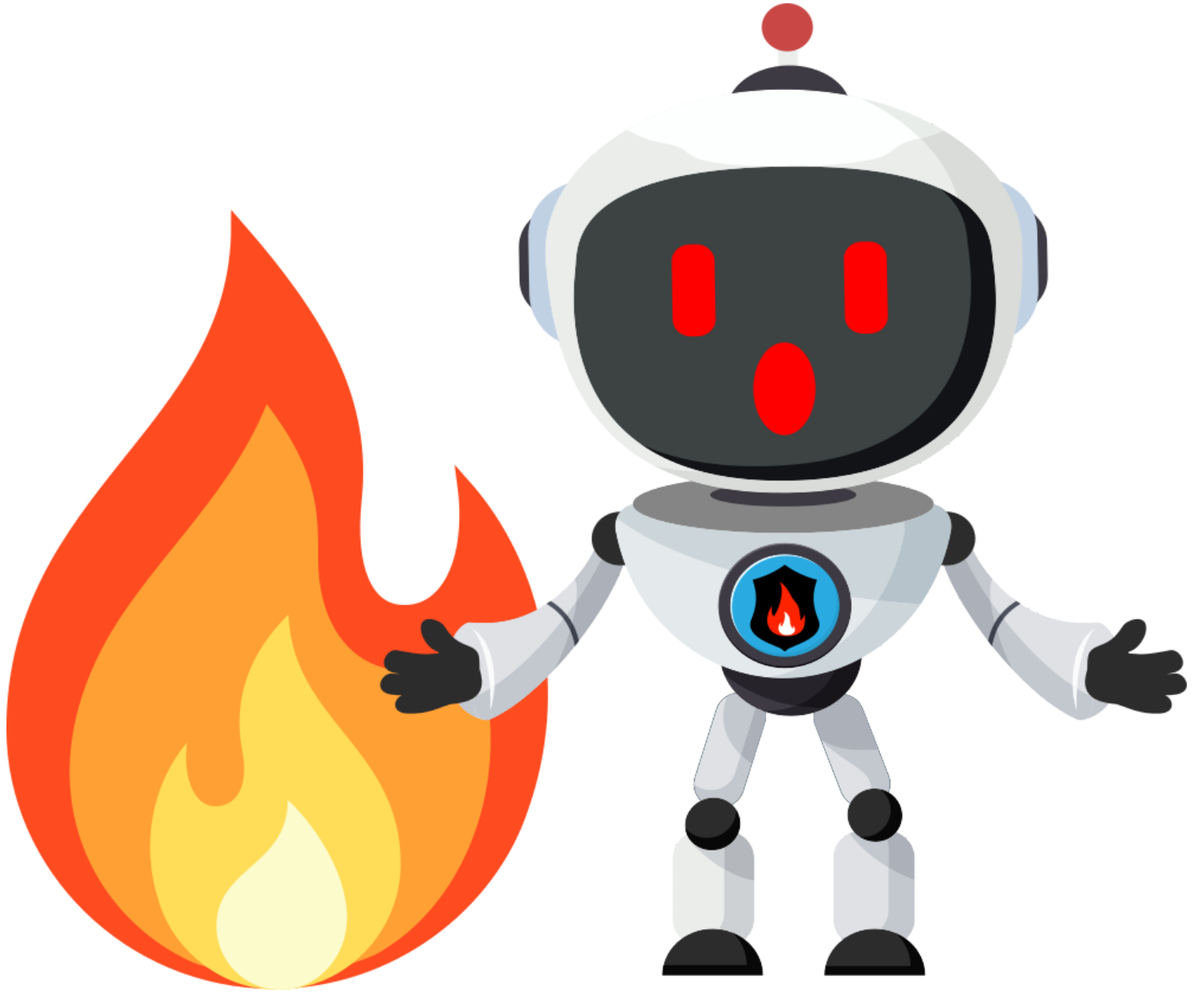


Accountability is vital

- People **trust** AI systems more when there is accountability
- Accountability ensures ethical use and mitigates potential harm
- Accountability means not absolving humans from responsibility

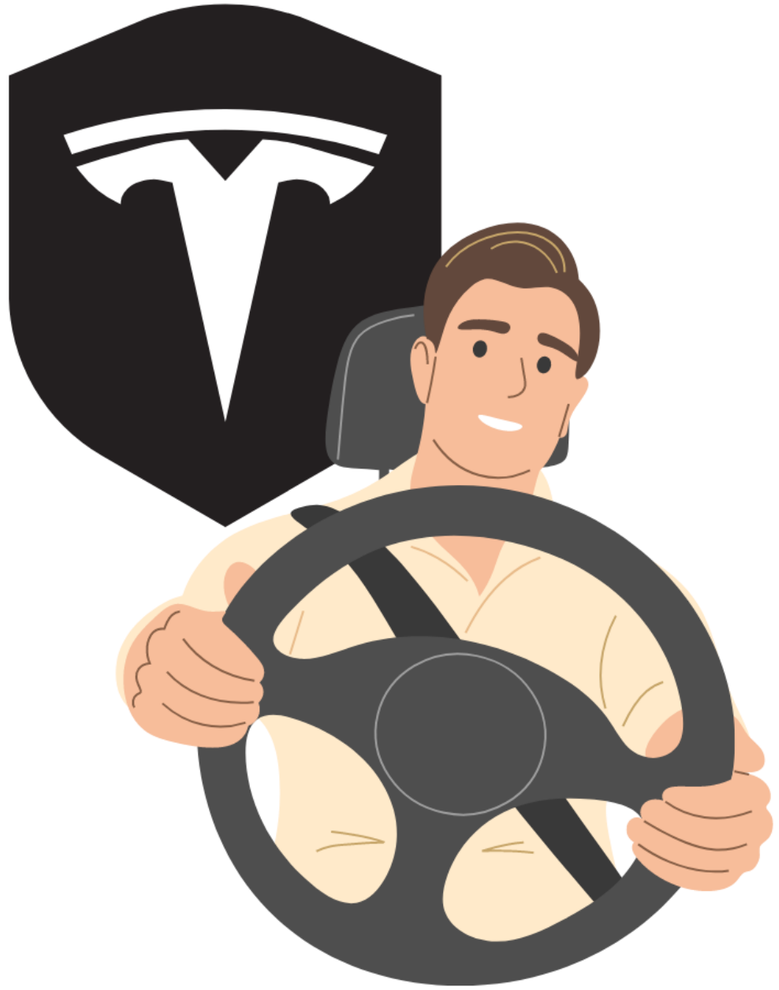
The paradox of accountability

- Increasing AI accountability can improve trust
- Yet, excessive trust in AI can lead to misguided decisions
- Example:
  - Georgia Tech study where participants followed misguided robot guidance



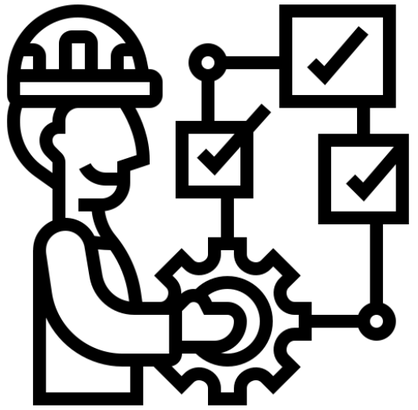
The Tesla story

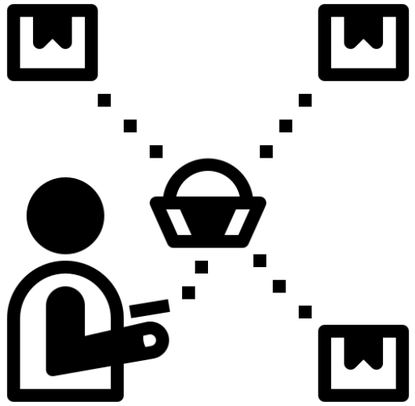
- Misunderstanding of the auto-pilot capabilities among consumers
- Criticism for Tesla's insufficient safeguards
- Both Tesla and consumers share responsibility



Achieving accountability

- AI producers:
  - Achieving accountability involves transparency and solving the 'Black Box' problem
  - Attributing responsibility is key



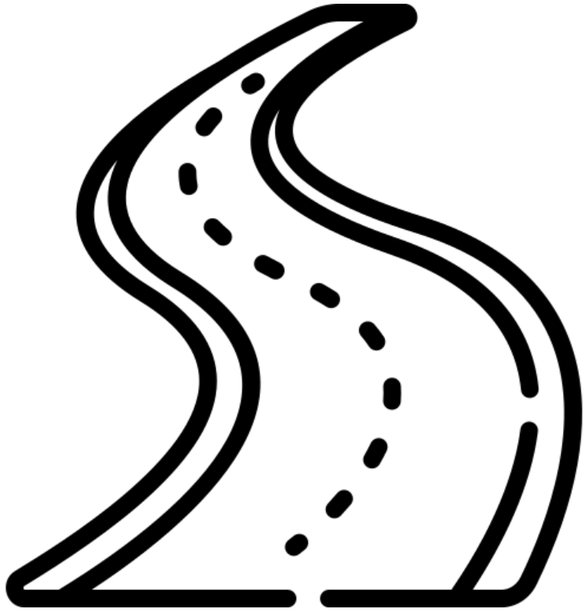


- AI consumers:
  - 'Trust but verify'
- Producers and consumers both play a role in creating ethical AI
- Challenges are opportunities for innovation

[1] Icons made by Eucalyp & Sumitsaengtong from www.flaticon.com

No one-size-fits-all

- Accountability in AI is a **continuous journey**
- With each AI advancement, the **accountability** conversation evolves
- No one-size-fits-all approach; varies across industries



[1] Icon made by Freepik from www.flaticon.com

Let's practice!

AI ETHICS

Explainable AI

AI ETHICS

Joe Franklin

Associate Data Literacy and Essentials Manager, DataCamp

datacamp

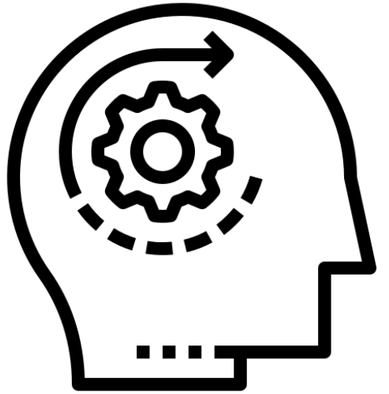
What's explainable AI?

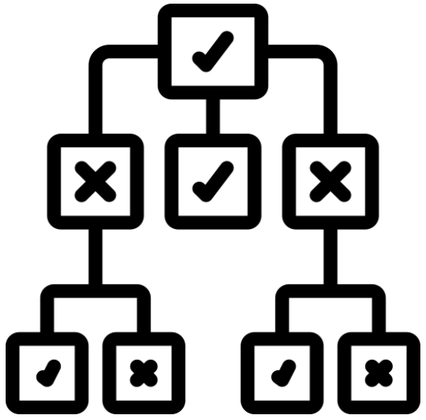
- AI systems whose internal workings are understood by humans
- Goal: Making AI's decision-making clear, understandable, and explainable
- Helps understand why and how AI makes decisions
- Major step towards ethical AI usage

[1] Icon made by vectorsmarket15 from www.flaticon.com

The central pillars

- **Transparency, fairness, accountability** are central
- AI conclusions should be accessible and logical to humans

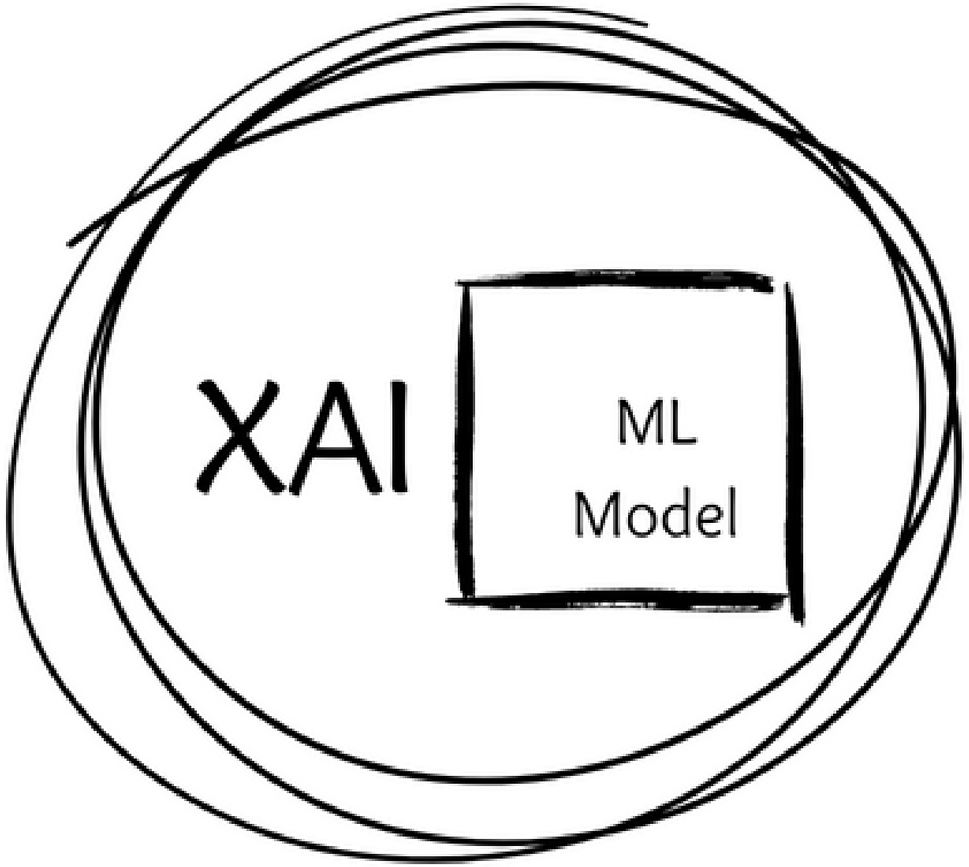




- Models built with explainability at their core
- Uses interpretable models like **decision trees** or **linear regression**
- Power in seeing the process, despite possibly lower performance

[1] Icons made by juicy_fish & Becris from www.flaticon.com

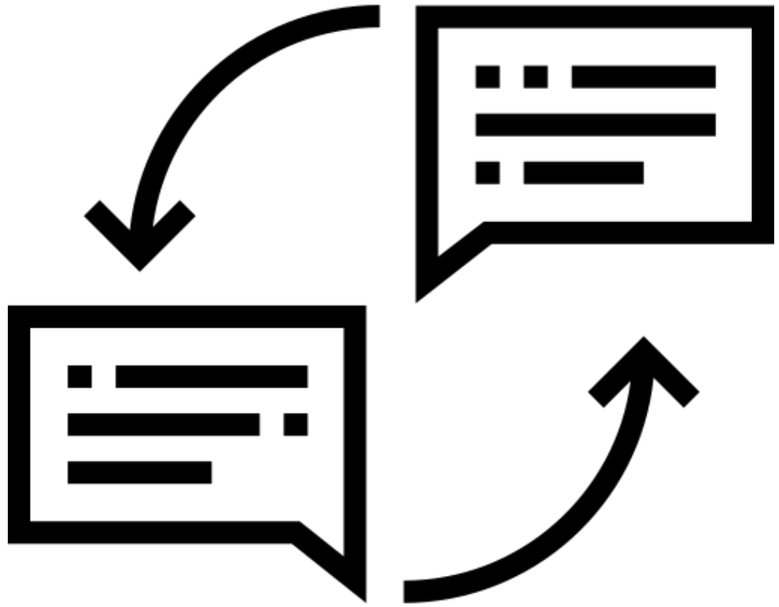
How does it work?



How does it work?

Local Interpretable Model-agnostic Explanations (LIME)

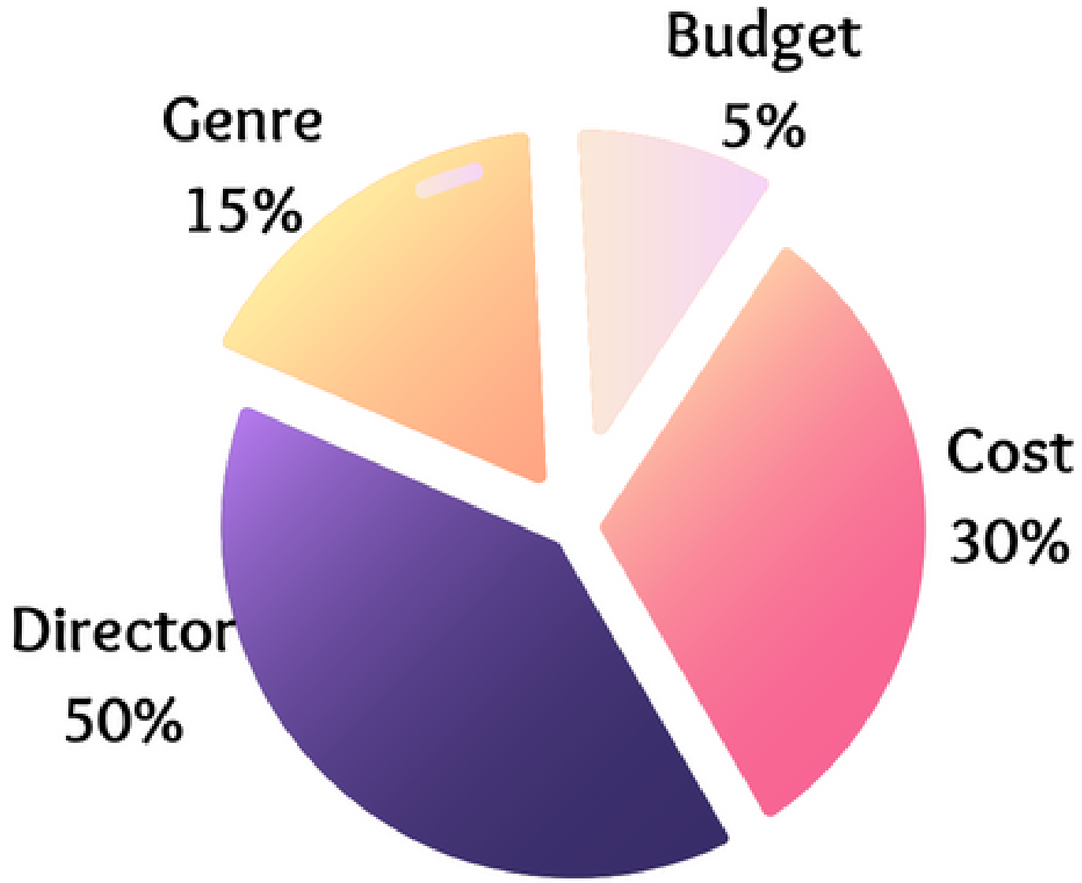
- LIME as a **translator** that helps the model **communicate**
- Creates a simpler version of the model's decision process for a specific prediction
- Example:
  - Explains a movie's hit prediction based on factors like director popularity and high budget



[1] Icon made by Freepik from www.flaticon.com

SHapley Additive exPlanations (SHAP)

- SHAP: A detective of AI, revealing **feature importance**
- SHAP in Action
  - Director: 50%
  - Cast: 30%
  - Genre: 15%
  - Budget: 5%



Future of XAI

- Many more **techniques** and **approaches** exist in XAI
- The gap between XAI and traditional AI is **shrinking**
- Ongoing research is improving AI interpretability

Let's practice!

AI ETHICS