Trustworthy and Socially Responsible AI

Transparent and eXplainable AI

Anahid Jalali

ETHICS AND AI POLICIES

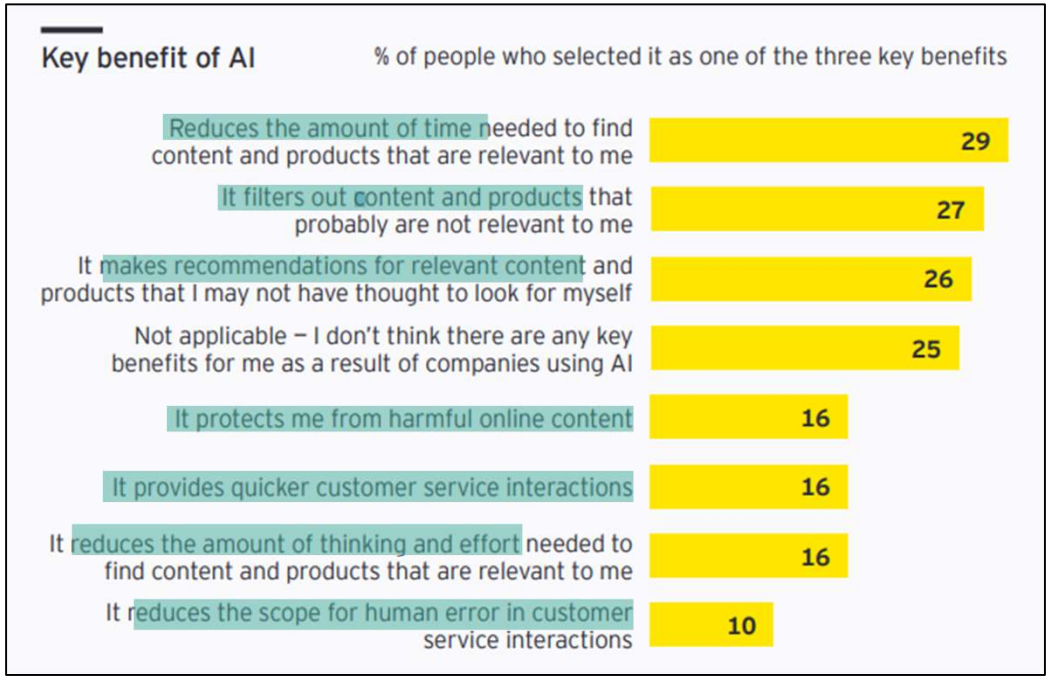
AUSTRIAN INSTITUTE OF TECHNOLOGY

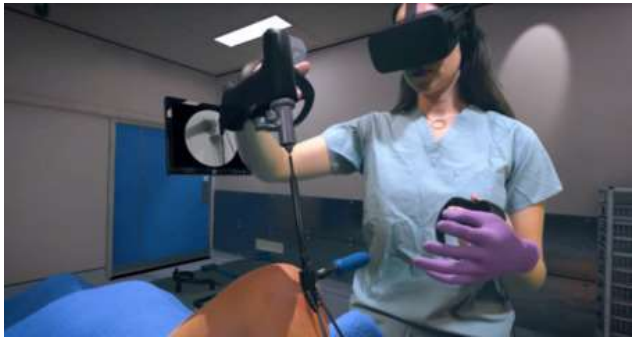
- **AI Policy Observatories and recommenders**
  - **Universal Guidelines for Artificial Intelligence**
  - **OECD AI Principles**
  - **UNESCO Recommendations**
  - **EU AI Act**
  - **COE AI Treaty**

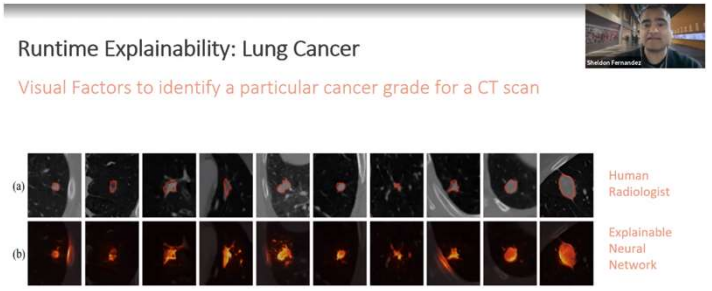
- **Defining redlines and limitations for safety and privacy**
- **Accountability**
- **Clear and transparent documentation of AI development**
- **Human centered values**
- **Sustainable and green**

https://www.archyworldys.com/missing-link-artificial-intelligence-justitias-friend-and-helper/

BENEFITS & RISKS OF AI

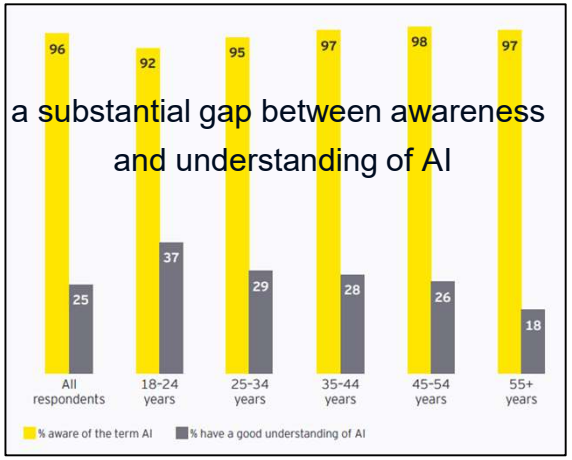


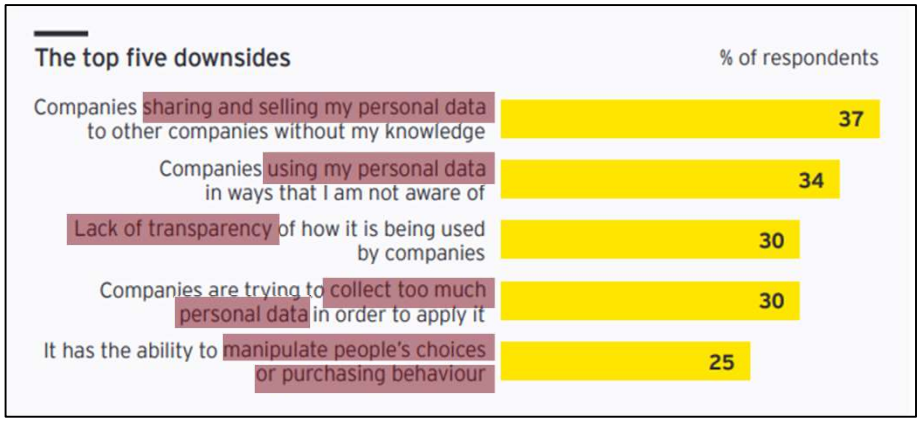




https://assets.ey.com/content/dam/ey-sites/ey-com/en_uk/topics/tmt/ey-ai-and-ethics.pdf

BENEFITS & RISKS OF AI

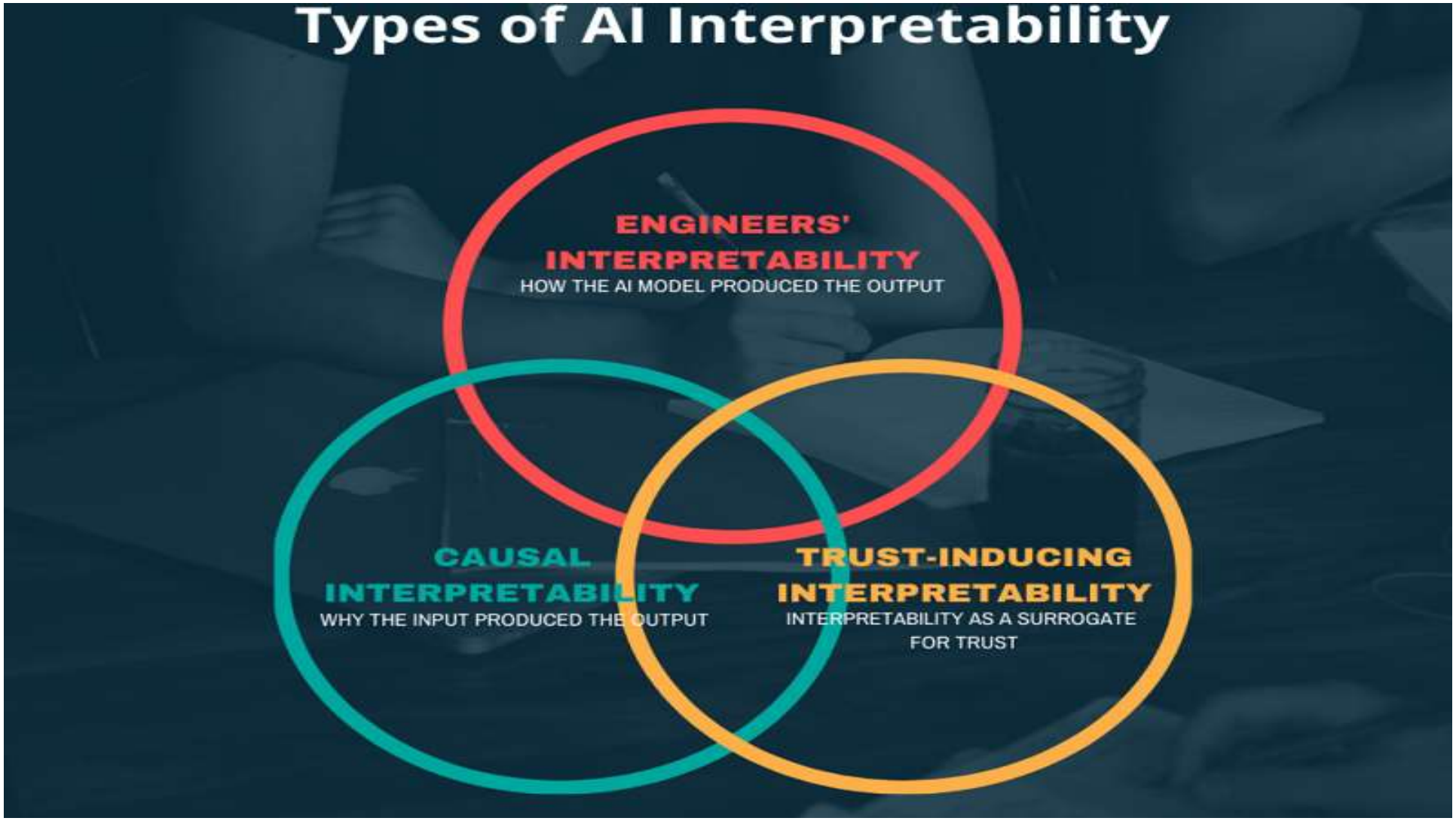




https://assets.ey.com/content/dam/ey-sites/ey-com/en_uk/topics/tmt/ey-ai-and-ethics.pdf

BENEFITS & RISKS OF AI

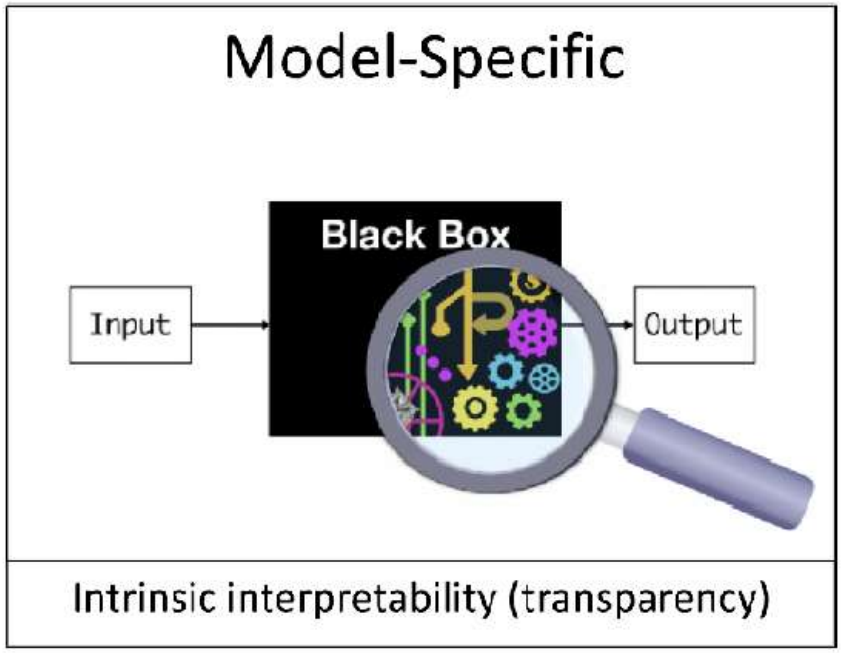
Majority of enterprise employees use non-approved SaaS to increase productivity	Lack of AI Implementation Traceability
AI systems from (potentially biased) historical data	Introducing Bias into Decision Making
Exponential growth of data increases the risks of exposing customer or employee data	Data Sourcing and Violation of Personal Privacy
Cannot explain why a banking customer is rejected based on the output of AI system	Black Box Algorithms and Lack of Transparency
In 2018, a self-driving car hit and killed a pedestrian. Who is legally responsible for the outcome? Is it the company, the programmer, or the system? Or the backup driver?	Unclear legal responsibilities

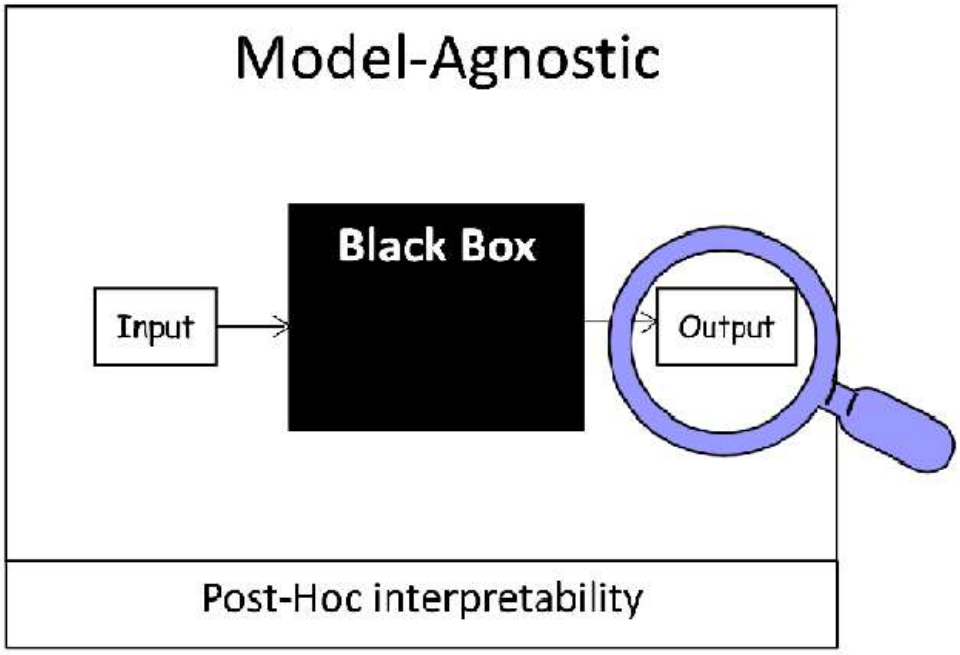

Types of AI Interpretability
ENGINEERS' INTERPRETABILITY
HOW THE AI MODEL PRODUCED THE OUTPUT
CAUSAL INTERPRETABILITY
WHY THE INPUT PRODUCED THE OUTPUT
TRUST-INDUCING INTERPRETABILITY
INTERPRETABILITY AS A SURROGATE FOR TRUST

EXPLAINABILITY/TRANSPARENCY/INTERPRETABILITY

AUSTRIAN INSTITUTE OF TECHNOLOGY

- Specific or Agnostic?





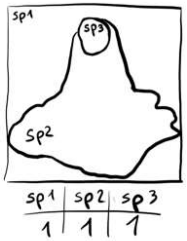
Ex-ante - data statistics, bias in data, definition of task, attributes used, scaling of attributes

EXPLAINABILITY/TRANSPARENCY/INTERPRETABILITY

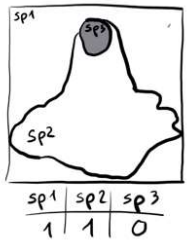
- SHAP

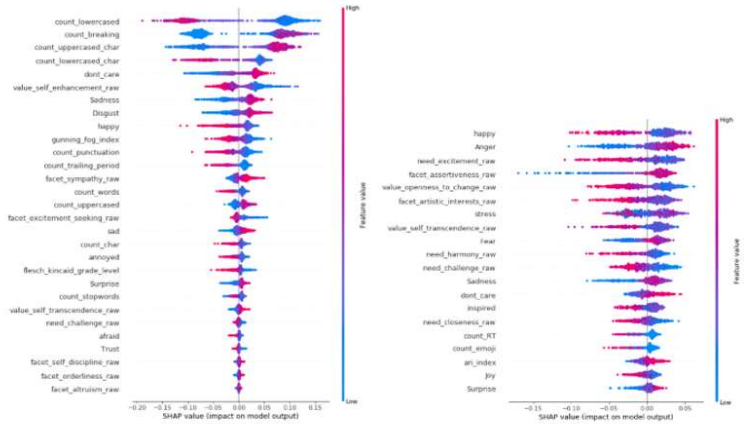
Coalitions of super pixels $\xrightarrow{h_x(z')}$ Image

Instance x

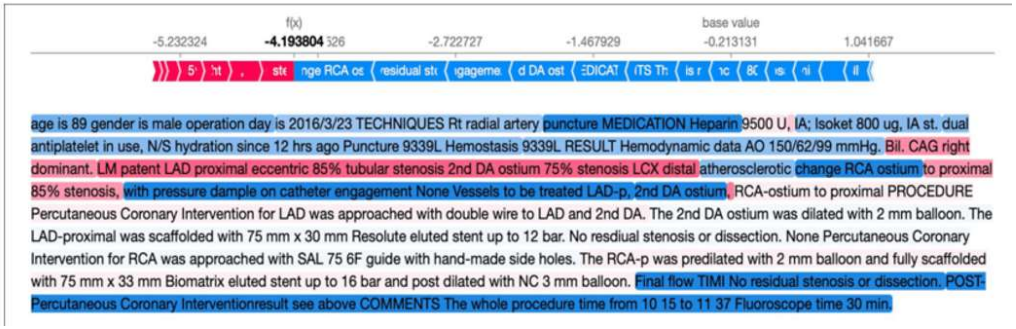


Instance x with absent features

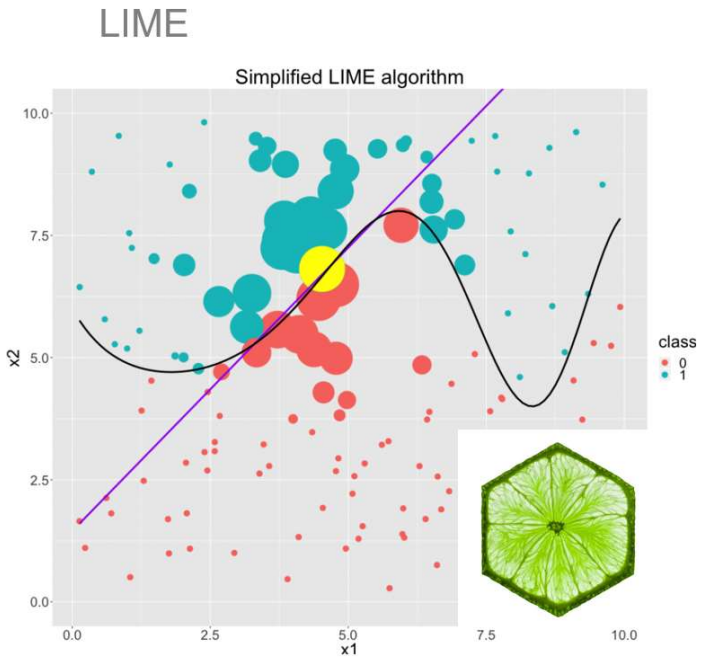


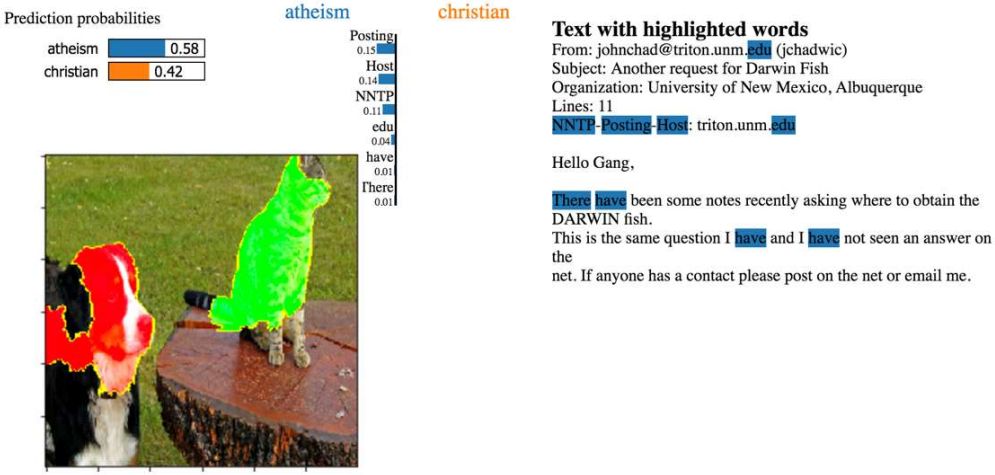


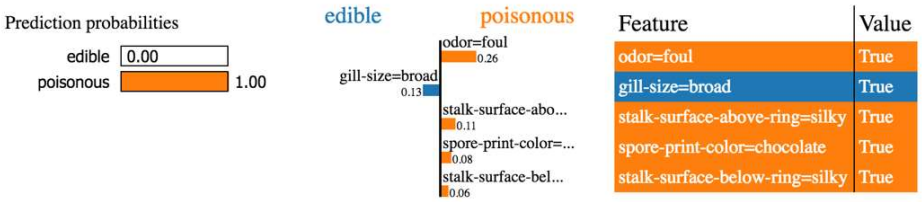
(a) PolitiFact (b) PAN



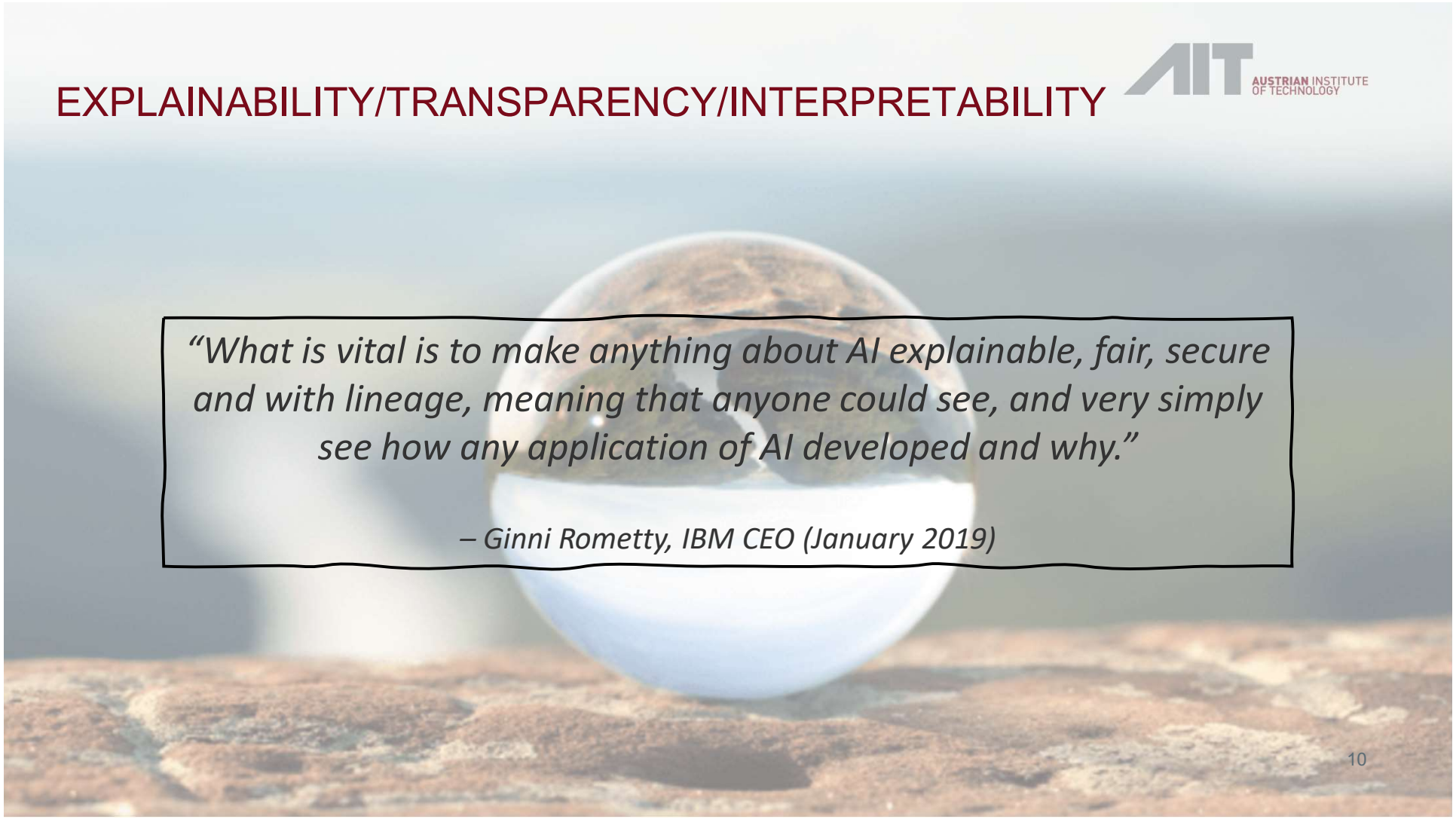
EXPLAINABILITY/TRANSPARENCY/INTERPRETABILITY







EXPLAINABILITY/TRANSPARENCY/INTERPRETABILITY



"What is vital is to make anything about AI explainable, fair, secure and with lineage, meaning that anyone could see, and very simply see how any application of AI developed and why."

– Ginni Rometty, IBM CEO (January 2019)

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

THANK YOU!

anahid.jalali@ait.ac.at