Tusting AI - Explainability

Harley Davis
VP, Automation Intelligence and France Lab
IBM
harley.davis@fr.ibm.com
@HarleyDavisPro
What do you trust?

Fuzzy Logic

Statistical Correlation

Machine Learning
Complex Algorithms
Who do you trust?

You: But why?
Loan officer: Our AI told us to.

You: How do you know it’s cancer?
Doctor: Our AI told us it was.
Elements of Trust

- Robustness
- Fairness
- **Explainability**
- Lineage

What’s the problem?

Machine Learning
Approaches to Explainability

- Correlation
- Causal Models
- Counterfactuals
- Decision Models
- Symbolic Rules
- Missing features

For people, explanations are symbolic, not mathematical
Who do you trust?

You: But why?
Loan officer: Your risk profile was good but your revenue to existing debt is too low for this loan type.

You: How do you know it’s cancer?
Doctor: The image of your liver showed several lesions that are strongly correlated with cancer.
Thank you!
Danke!
Gracias!
Grazie!
Merci!
Obrigado!
And so on...

EU AI Ethics Guidelines

Trustworthy AI should respect all applicable laws and regulations, as well as a series of requirements; specific assessment lists aim to help verify the application of each of the key requirements:

• **Human agency and oversight**: AI systems should enable equitable societies by supporting human agency and fundamental rights, and not decrease, limit or misguide human autonomy.

• **Robustness and safety**: Trustworthy AI requires algorithms to be secure, reliable and robust enough to deal with errors or inconsistencies during all life cycle phases of AI systems.

• **Privacy and data governance**: Citizens should have full control over their own data, while data concerning them will not be used to harm or discriminate against them.

• **Transparency**: The traceability of AI systems should be ensured.

• **Diversity, non-discrimination and fairness**: AI systems should consider the whole range of human abilities, skills and requirements, and ensure accessibility.

• **Societal and environmental well-being**: AI systems should be used to enhance positive social change and enhance sustainability and ecological responsibility.

• **Accountability**: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes.