Transparency AI

- What is Explainable AI or XAI?

  artificial intelligence in which the results of the solution can be understood by humans

- What is the alignment problem?

  how to ensure models capture our norms and values, understand what we mean or intend, and do what we want

- What is a rule-based model?

  an easily interpreted machine learning system

  they take the form of a list of "if x then y" rules

- Why is transparency necessary?

  a model could learn the wrong rules leading to flawed decision making

- What is an example of a wrong rule?

  Wrong rule: If a patient has a history of asthma, then they are low-risk for pneumonia and you should treat them as an outpatient.

  Why is this wrong? Asthma is a serious risk factor for pneumonia so they are put right into critical care units.

  Essentially: a pattern was observed in the data that was not interpreted correctly.

- What is a generalized additive model?

  a collection of graphs, each of which represents the influence of a single variable, e.g. a graph may show risk as a function of age

- What is the benefit of a generalized additive model?

  you can visualize on a plain 2D graph every factor going into the model, making it easier to identify any strange patterns

- General Data Protection Regulation (EU) in 2018

  people have the right to ask for an explanation of algorithmically made decisions

-

-

-

-

-

-

-