Opacity and the black-box problem in AI

- In which sense is a model opaque?
- XAI = explainable AI
- Different responses to opacity?
  - Interpretability
  - Explainability
  - Accountability
  - Intelligibility
  - Understandability

Is the black-box problem really a problem?

- Business/deployment concerns
  - Building trust to secure usage
  - Informativeness/usefulness of the models for the end-user
  - Transferability/generalizability
- Ethical concerns
  - Biases, fairness
  - Privacy
- Legal requirements
  - GDRP in Europe (having access to explanations for decisions taken by algorithms)
  - Liability concerns (attributing liability in case of dysfunctions), accountability
- Improvement/efficiency concerns
  - Understanding why a model works and doesn't work (dysfunction)
- Curiosity/scientific knowledge about the functioning of AI
  - Cf neurosciences and the 'brain black box'
- Scientific understanding
  - Getting at causality/explanation of natural phenomena when modeled by AI

Answering the black-box problem

- A. Transparency solutions?
  - Simulatability/mental model
  - Interpretable models (from start)
  - Decomposability
  - Algorithmic transparency…
- B. Justification/Post-hoc interpretability?
  - Feature based explanation
    - Justificative (points to specific evidence in the inputs)
    - Ontological or formal (points to essential properties of the object to be classified)
    - Explanation by example
  - Introspective
    - By looking at the neurons, explaining how the network came to the classification output
  - Interventionist
    - Feeding the model with specific sets of (simplified) inputs so as to see which types of outputs follow
  - Model approximation/local explanation …

Evaluating XAI solutions as explanations

- In which sense are the solutions proposed “good explanations”?
  - Contextual utility of the explanation (for the end-user)
  - Comparison with formal/philosophical models of explanation
  - Explanation or understanding? Or justification?
  - Source of explanatory power?