# Probing Further into “Interpretability”: Caveats & Challenges

# Agenda

- Two Papers:
  - Transparency: Motivations and Challenges
  - The Mythos of Model Interpretability

- Break out groups

- Discussion

# The Mythos of Model Interpretability

Zachary Lipton; 2017

# Contributions

- Goal: Refine the discourse on interpretability
- Outline desiderata of interpretability research
  - Motivations for interpretability are often diverse and discordant
- Identifying model properties and techniques thought to confer interpretability

# Motivation

- We want models to be not only good w.r.t. predictive capabilities, but also interpretable

- Interpretation is underspecified
  - Lack of a formal technical meaning

- Papers provide diverse and non-overlapping motivations for interpretability

# Prior Work: Motivations for Interpretability

Interpretability promotes trust

- But what is trust?
- Is it faith in model performance?
- If so, why are accuracy and other standard performance evaluation techniques inadequate?

# When is interpretability needed?

- Simplified optimization objectives fail to capture complex real life goals.
  - Algorithm for hiring decisions – productivity and ethics
  - Ethics is hard to formulate

- Training data is not representative of deployment environment

Interpretability serves those objectives that we deem important but struggle to model formally!

# Desiderata

- Understanding motivations for interpretability through the lens of prior literature
  - Trust
  - Causality
  - Informativeness
  - Fair and Ethical Decision Making

# Desiderata: Trust

- Is trust simply confidence that the model will perform well?

- A person might feel at ease with a well understood model, even if this understanding has no purpose

- Training and deployment objectives diverge
  - E.g., model makes accurate predictions but not validated for racial biases

- Trust ▭ relinquish control
  - For which examples is the model right?

# Desiderata: Causality

- Researchers hope to infer properties (beyond correlational associations) from interpretations/explanations
  - Regression reveals strong association between smoking and lung cancer

- However, task of inferring causal relationships from observational data is a field in itself
  - Don Rubin
  - Judea Pearl

# Desiderata: Informativeness

- Predictions □ Decisions
  - Convey additional information to human decision makers

- Example: Which conference should I target?
  - A one word answer is not very meaningful

- Interpretation might be meaningful even if it does not shed light on model’s inner workings
  - Similar cases for a doctor in support of a

# Desiderata: Fair & Ethical Decision Making

- ML is being deployed in critical settings
  - Eg., healthcare

- How can we be sure algorithms do not discriminate on the basis of race?
  - AUC is not good enough

- Side note: European Union – Right to explanation

# Properties of Interpretable Models

- Transparency
  - How exactly does the model work?
  - Details about its inner workings, parameters etc.

- Post-hoc explanations:
  - What else can the model tell me?
  - Eg., visualizations of learned model, explaining by example

# Transparency: Simulatability

- Can a person contemplate the entire model at once?
  - Need a very simple model

- A human should be able to take input data and model parameters and calculate prediction

# Transparency: Decomposability

- Understanding each input, parameter, calculation
  - E.g., decision trees, linear regression

- Inputs must be interpretable
  - Models with highly engineered or anonymous features are not decomposable

# Algorithmic Transparency

- Learning algorithm itself is transparent
  - E.g., linear models (error surface, unique solution)

- Modern deep learning methods lack this kind of transparency
  - We don’t understand how the optimization methods work
  - No guarantees of working on new problems

- **Note:** Humans do not exhibit any of these forms of transparency

# Post-hoc: Text Explanations

- Humans often justify decisions verbally (post-hoc)

- Krening et. al.:
  - One model is a reinforcement learner
  - Another model maps models states onto verbal explanations
  - Explanations are trained to maximize likelihood of ground truth explanations from human players
  - So, explanations do not faithfully describe agent decisions, but rather human intuition

# Post-hoc: Visualization

- Visualize high-dimensional data with t-SNE
    - 2D visualizations in which nearby data points appear close

- Perturb input data to enhance activations of certain nodes in neural nets (image classification)
    - Helps understand which nodes corresponds to what aspects of the image
    - Eg., certain nodes might correspond to dog faces

# Post-hoc: Example Explanations

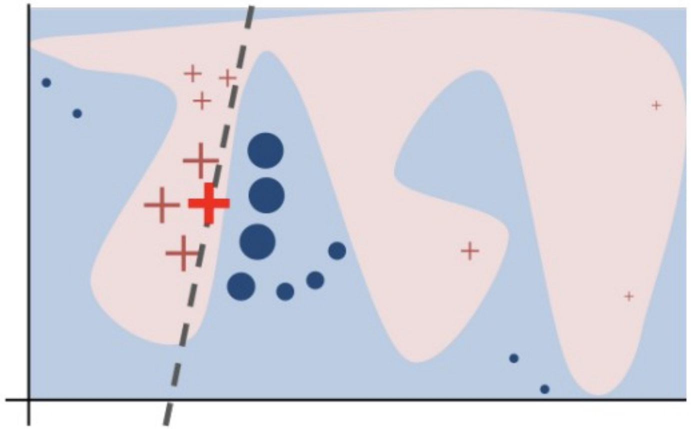
- Reasoning with examples

- E.g., Patient A has a tumor because he is similar to these k other data points with tumors

- k neighbors can be computed by using some distance metric on learned representations
  - Eg., word2vec

# Post-hoc: Local Explanations

- Hard to explain a complex model in its entirety
  - How about explaining smaller regions?



LIME (Ribeiro et. al.)

  - Explains decisions of any model in a local region around a particular point

  - Learns sparse linear model

# Claims about interpretability must be qualified

- If a model satisfies a form of transparency, highlight that clearly

- For post-hoc interpretability, fix a clear objective and demonstrate evidence

# Transparency may be at odds with broader objectives of AI

- Choosing interpretable models over accurate ones to convince decision makers

- Short term goal of building trust with doctors might clash with long term goal of improving health care

# Post-hoc interpretations can mislead

- Do not blindly embrace post-hoc explanations!

- Post-hoc explanations can seem plausible but be misleading
  - They do not claim to open up the black-box;
  - They only provide plausible explanations for its behavior
  - E.g., text explanations

# Summary

- Goal: Refine the discourse on interpretability

- Outline desiderata of interpretability research
  - Motivations for interpretability are often diverse and discordant

- Identifying model properties and techniques thought to confer interpretability

# Transparency: Challenges and Motivation

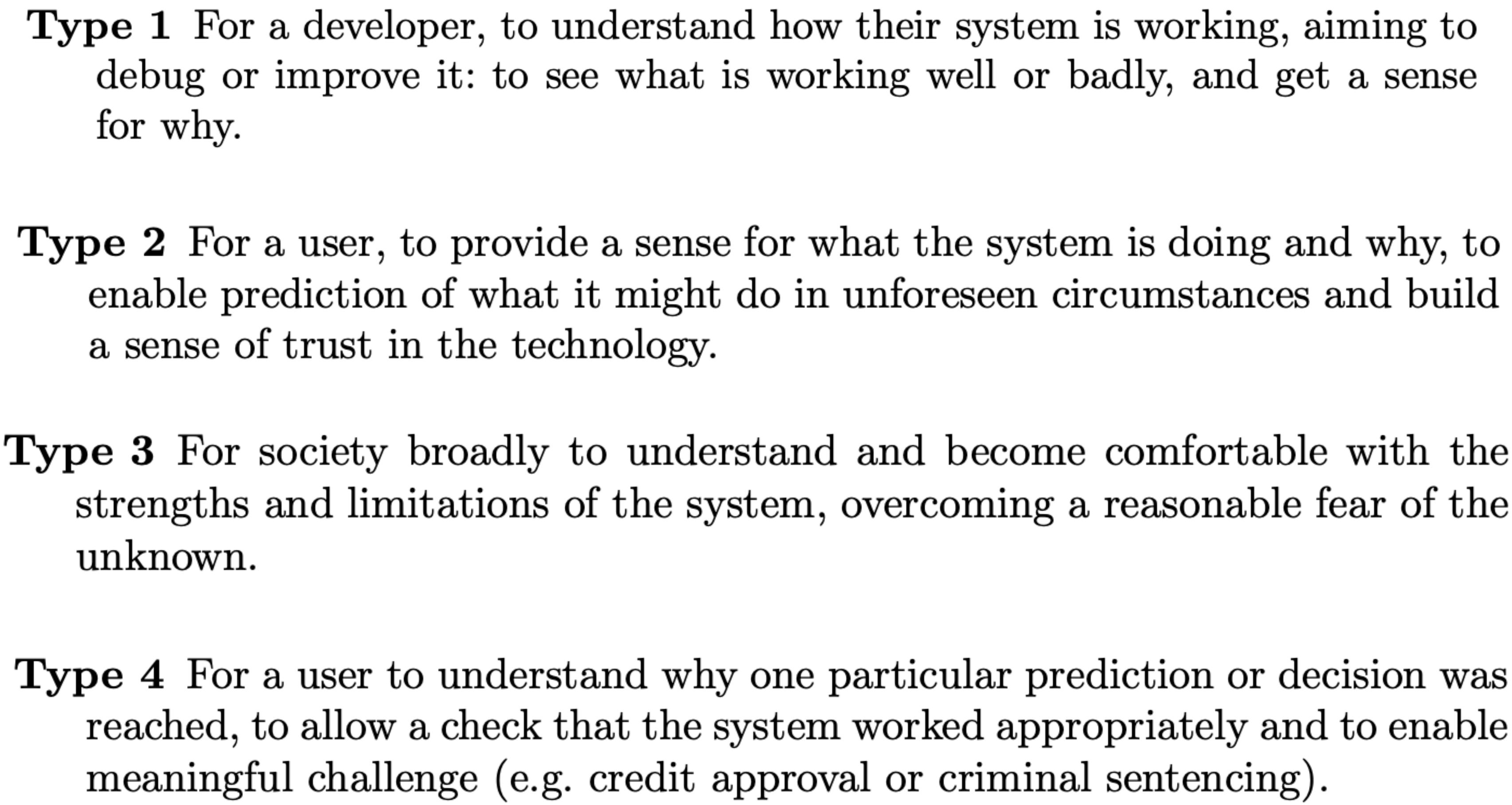
Adrian Weller; 2019

# Contributions

- Characterizing different kinds of transparency, and underlying motivations

- Shedding light on the downsides of having transparency

# Types and Goals of Transparency



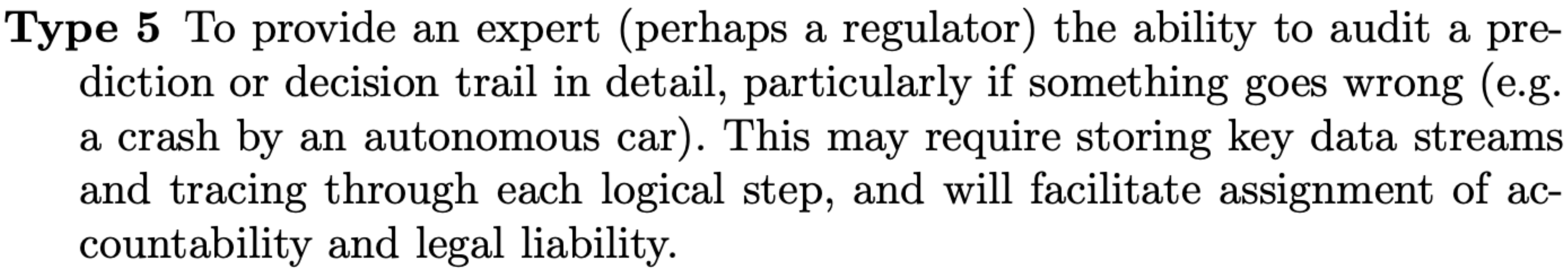
**Type 1** For a developer, to understand how their system is working, aiming to debug or improve it: to see what is working well or badly, and get a sense for why.

**Type 2** For a user, to provide a sense for what the system is doing and why, to enable prediction of what it might do in unforeseen circumstances and build a sense of trust in the technology.

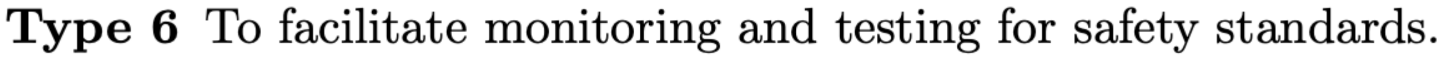
**Type 3** For society broadly to understand and become comfortable with the strengths and limitations of the system, overcoming a reasonable fear of the unknown.

**Type 4** For a user to understand why one particular prediction or decision was reached, to allow a check that the system worked appropriately and to enable meaningful challenge (e.g. credit approval or criminal sentencing).

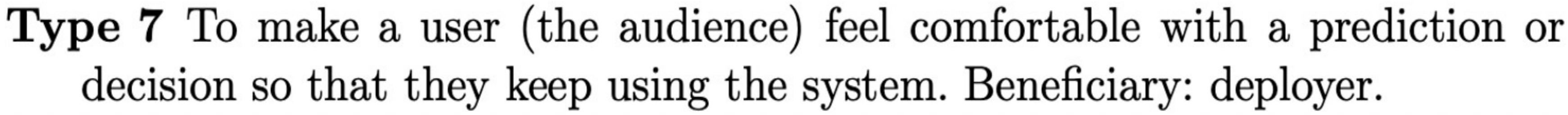
# Types and Goals of Transparency



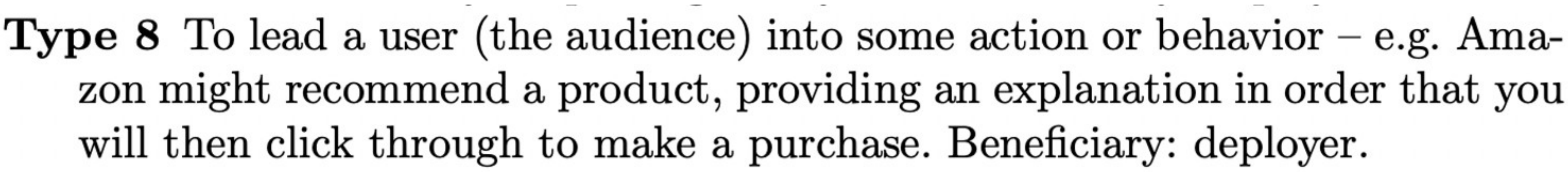
**Type 5** To provide an expert (perhaps a regulator) the ability to audit a prediction or decision trail in detail, particularly if something goes wrong (e.g. a crash by an autonomous car). This may require storing key data streams and tracing through each logical step, and will facilitate assignment of accountability and legal liability.



**Type 6** To facilitate monitoring and testing for safety standards.



**Type 7** To make a user (the audience) feel comfortable with a prediction or decision so that they keep using the system. Beneficiary: deployer.



**Type 8** To lead a user (the audience) into some action or behavior – e.g. Amazon might recommend a product, providing an explanation in order that you will then click through to make a purchase. Beneficiary: deployer.

# Global vs. Local

- Global: understanding whole system
  - Types 2 – 3

- Local: explanation for a particular prediction
  - Types 4, 5, 7, 8

# Types and Goals of Transparency

- Explanations are beneficial to the society only if they are faithful

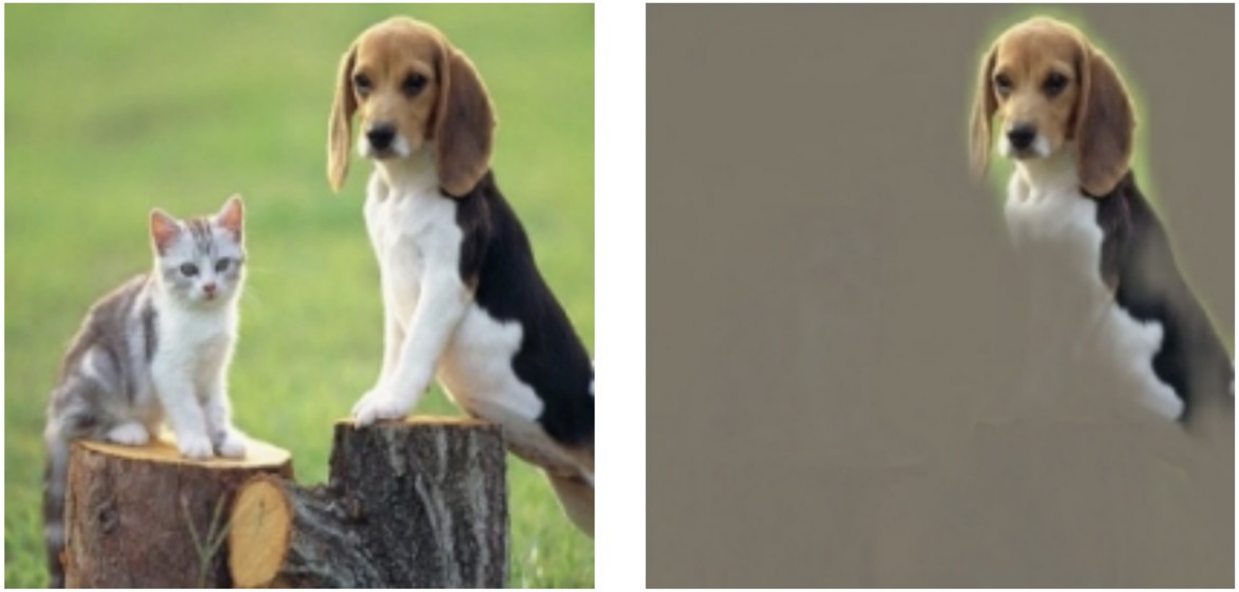
- Notion of faithfulness is hard to characterize precisely!

- Defining criteria and tests for practical faithfulness are important open problems
  - Context is important!

# Types and Goals of Transparency

- Another challenge: Is an explanation good at conveying faithful information in understandable form, and if a human has actually understood it well?
- Context dependent
- Need for deeper probing!

# Comparing explanations is also hard!



If we have two different saliency maps, how to know which one is better

# Possible Dangers: Audience vs. Beneficiary

- Recommender systems
  - Amazon (Beneficiary) and its Users (Audience)

- Healthcare
  - Google Verily (Beneficiary) and Hospitals/Doctors (Audience)

- Criminal Justice
  - COMPAS (Beneficiary) and Courts/Judges (Audience)

# Government Use of Algorithms

- COMPAS system predicts risk of recidivism

- A prisoner should have some transparency into the decision made by COMPAS
    - To ensure proper process has been followed
    - Enable potential challenge

- But, can there be too much transparency?

- Also, recent push for making all code/data for such models public. Good idea?

# Gaming, IP Incentives, and Privacy

- If all details available, the process can be gamed
- Less incentive for private IP and slow progress
- Privacy and transparency are often in conflict
    - How much transparency is too much in a setting?

# Means and Ends

- Transparency □ reliability, fairness
- If we are able to develop a good set of “safety checks”, then may be it is ok to not have full transparency
    - E.g., autonomous vehicles

# Does giving more information to each individual help society?

- Braess’ paradox: more information empowers the agents to optimize their own agendas more efficiently, and thus may lead to a worse global outcome

# Selective Transparency & Discrimination

- Selective Transparency may hurt people disproportionately

- Furthermore, transparency about certain attributes (e.g., gender) in certain settings is known to cause discriminatory behavior as well

# Break Out Groups

Based on the different papers we have covered, can you summarize your stance on:

- When/Why is interpretability needed?
- When is interpretability a bad idea?
- Are there any privacy concerns around making models interpretable? What about fairness concerns?
- In real world settings, there have been cases where simpler models were chosen over accurate ones to secure “trust” of  decision makers. What do you think about this?