ARTIFICIAL INTELLIGENCE

WHY EXPLANATIONS MATTER

Albert Weichselbraun
University of Applied Sciences of the Grisons

PDF

AGENDA

1. Why explanations matter
2. Explainable Artificial Intelligence
3. Conclusions

WHY EXPLANATIONS MATTER

- Raise awareness of limitations
- Ethics and accountability

RAISE AWARENESS OF LIMITATIONS

Biases

Implicit biases in training data.

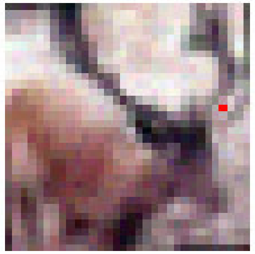
queen ~ king	**sister ~ brother**	**mother ~ father**
waitress ~ waiter	ovarian cancer ~ prostate cancer	convent ~ monastery
nurse ~ surgeon	registered nurse ~ physician	housewife ~ shopkeeper
giggle ~ chuckle	interior designer ~ architect	charming ~ affable
volleyball ~ football	cosmetics ~ pharmaceuticals	diva ~ superstar

(Examples from Bolukbasi et al, 2016)

FH GR

RAISE AWARENESS OF LIMITATIONS

Issues: Benign Conditions



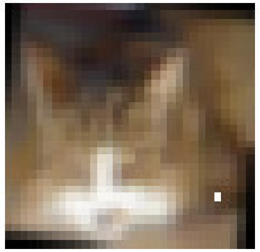
DEER
AIRPLANE(49.8%)

BIRD
FROG(88.8%)

SHIP
AIRPLANE(88.2%)

HORSE
DOG(88.0%)

SHIP
AIRPLANE(62.7%)



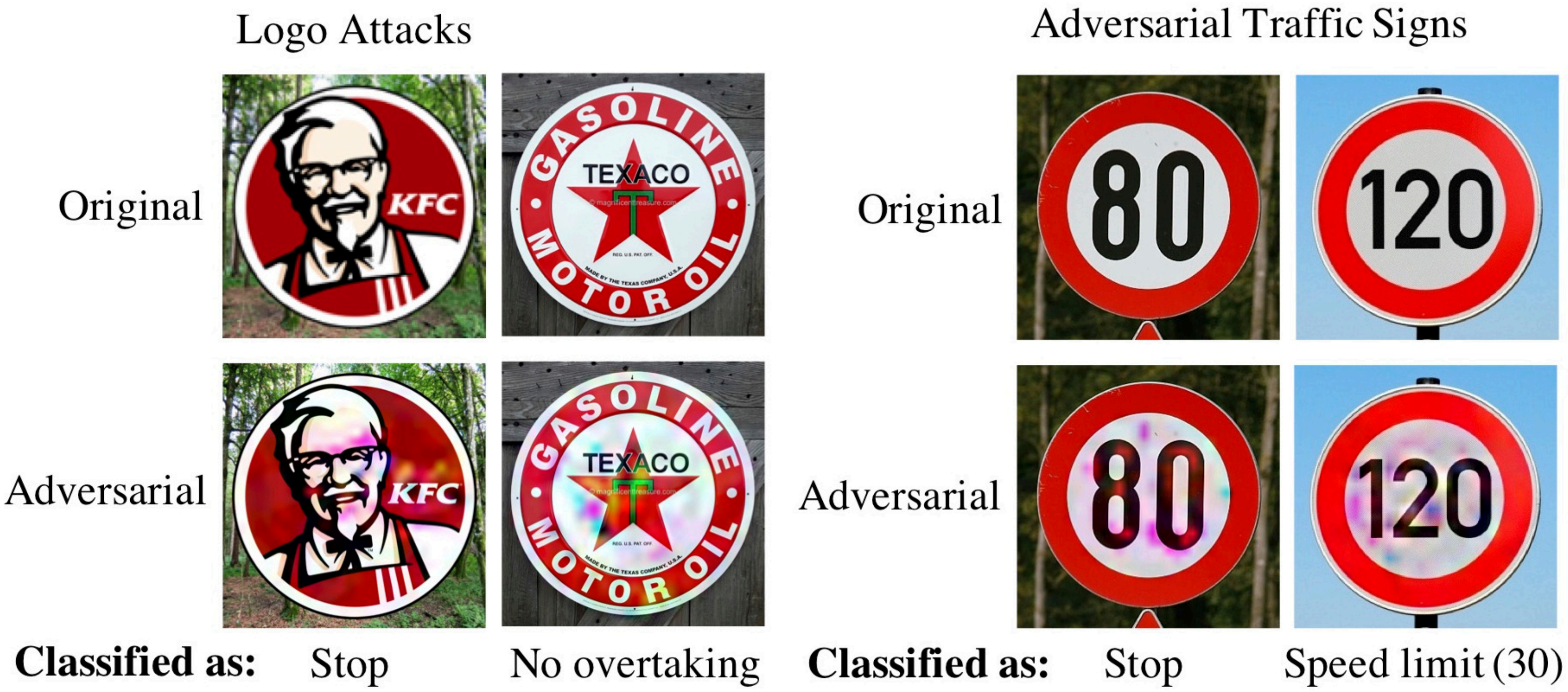
CAT
DOG(78.2%)

Jellyfish
Bathing tub(21.18%)

(Source: Su et al., 2018)

RAISE AWARENESS OF LIMITATIONS

Issues: Benign Conditions



(Source: Sitawarin et al, 2018)

ETHICS AND ACCOUNTABILITY

Fundamental principles relevant to Artificial Intelligence

1. Explainability
2. Justice
3. Non-maleficence
4. Autonomy

ETHICS AND ACCOUNTABILITY

Predictive sentencing (Starr, 2013)

Predictive sentencing involves a prediction of the risk or threat to society by the offenders and of the reaction of different types of offenders to different types of treatment modalities.

- COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)
- Accuracy: 0.71
- Algorithm uses features such as poverty, postal codes and employment status → highly correlated with minorities

FH
GR

ETHICS AND ACCOUNTABILITY

What is fair?

- literature defines many different kinds of fairness
  - e.g., group unaware, group thresholds, demographic parity, equal opportunity, equal accuracy, etc.
- first step: determine the "kind of fairness" you aim for
  - this is a decision that needs to be made by humans
  - determines the design goals for the AI
  - additional information: Google What-if-tool - AI Fairness

FH
GR

EXPLAINABLE ARTIFICIAL INTELLIGENCE

1. Explainability versus interpretability
2. Approaches and limitations

EXPLAINABILITY VERSUS INTERPRETABILITY

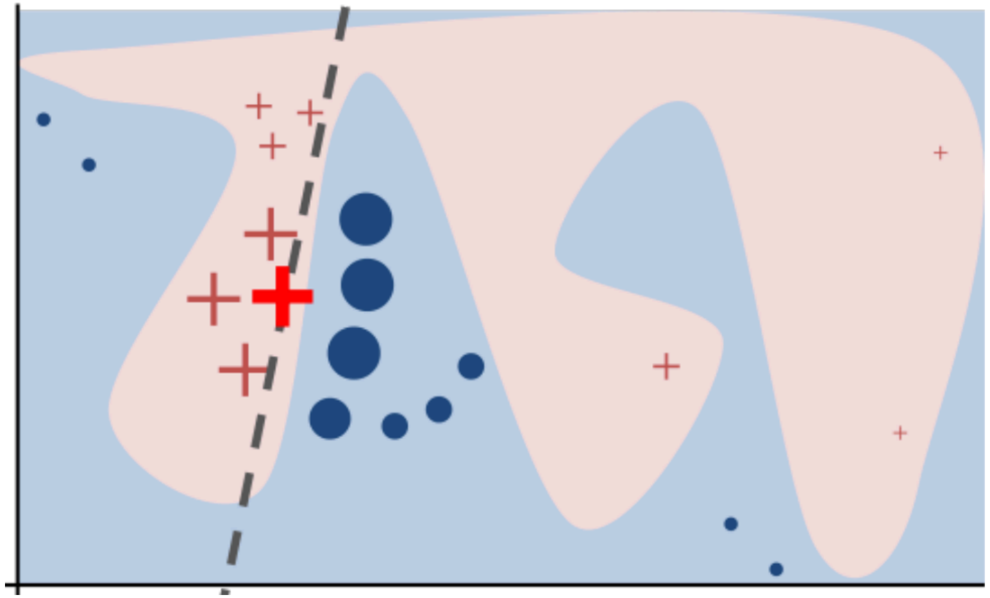
- **Interpretable models:** can be understood by humans without any other aids/methods
  $\rightarrow$ examples: linear regression (two or three model parameters), decision trees, symbolic AI
- **Explainable models:** need additional techniques to be "understood" by human (post-hoc explanations)
  $\rightarrow$ GPT-3 - model 175 billion parameters

APPROACHES AND LIMITATIONS

Post-hoc Explanations

Local Interpretable Model-Agnostic Explanations (LIME)

- explain singular predictions
- instability (explanations may vary between runs)



Local Interpretable Model-Agnostic Explanations
(Source: Ribeiro et al., 2016)

FH GR

APPROACHES AND LIMITATIONS

Post-hoc Explanations

Local Interpretable Model-Agnostic Explanations (LIME)

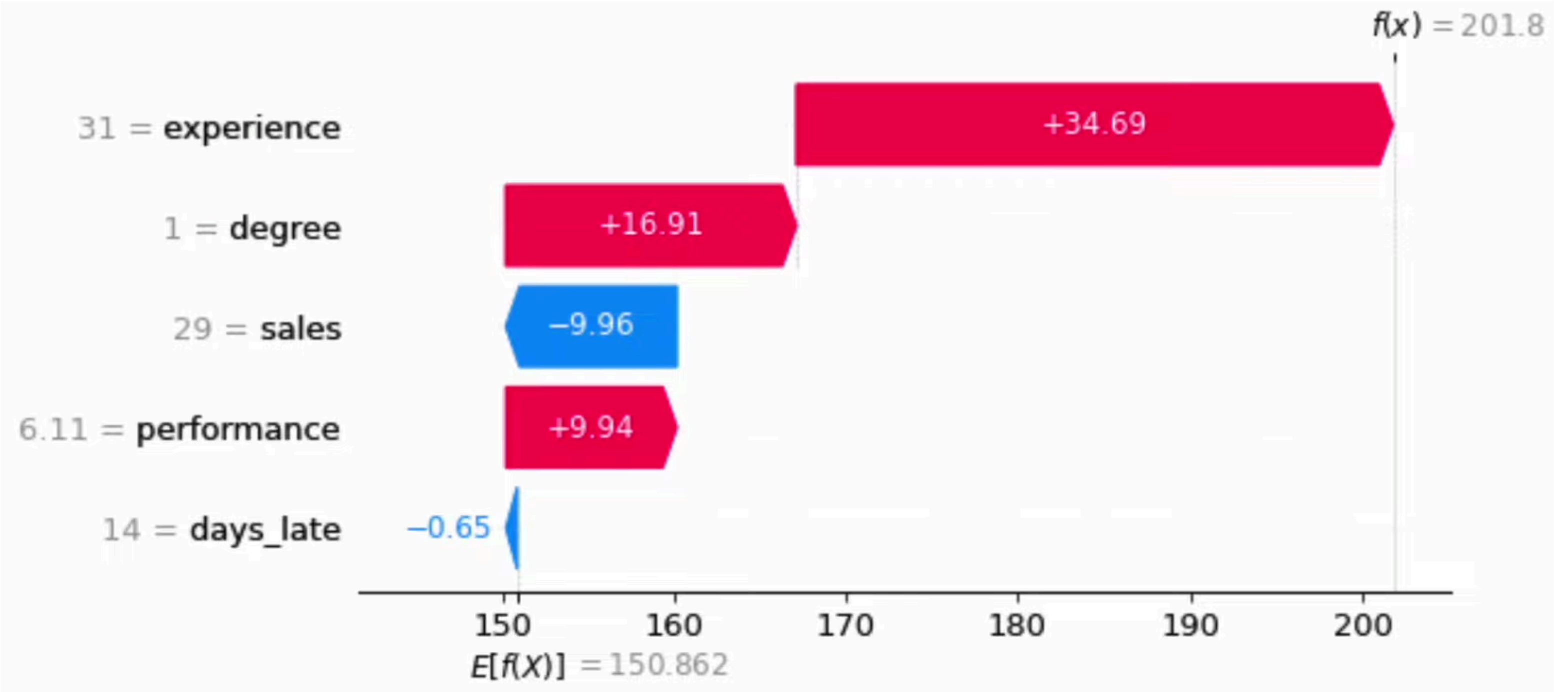
(a) Husky classified as wolf (b) Explanation

(Source: Ribeiro et al., 2016)

Post-hoc Explanations

SHAP (SHapley Additive exPlanations)

- explain singular predictions
- illustrate the contribution of each feature to the overall result
- reliability issues (SHAP values)



(Source: SHAP - A Data Odyseey)

CONCLUSIONS

CONCLUSIONS

- Why explanations matter?
  - help in understanding model limitations
  - provide insights into ethical issues
- Explainable AI
  - models are too large to be interpretable
  - post-hoc explanations $\rightarrow$ issues with reliability and helpfulness
  - still a major research challenge

FH GR