On the challenges of bringing explainable AI to practice

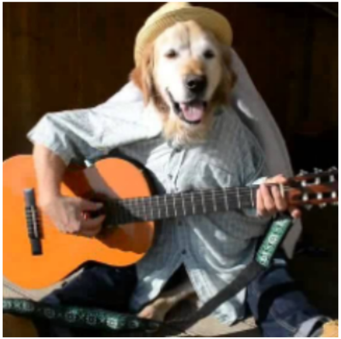
ADS meets CIDR 2020

Hinda Haned

January 15th, 2020

h.haned@uva.nl

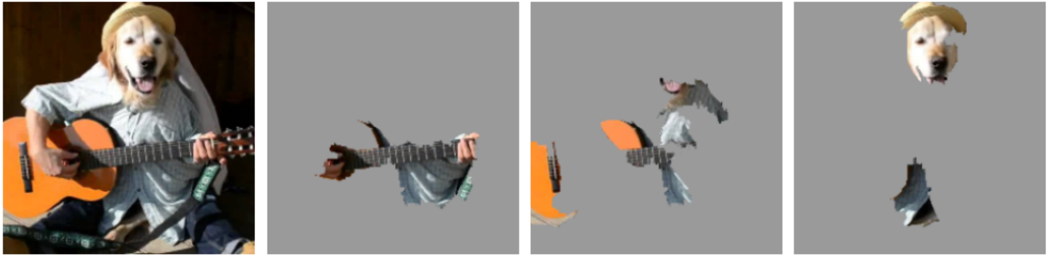
Example: classification task



Top predicted classes: Electric guitar, Acoustic guitar, Labrador

Ribeiro et al. Why Should I Trust You?: Explaining the Predictions of Any Classifier. KDD 2016.

Example: explaining predictions

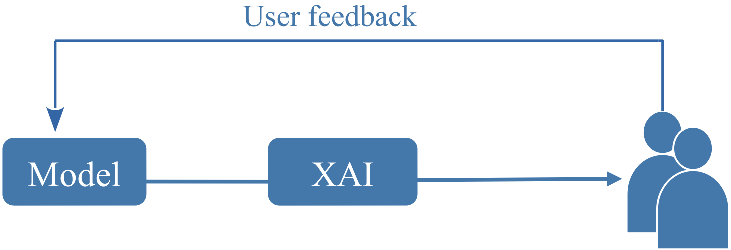


(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Ribeiro et al. Why Should I Trust You?: Explaining the Predictions of Any Classifier. KDD 2016.

XAI: eXplainable AI

Model outputs must be understandable and transparent to the decision makers and the subjects impacted by them



Explainable vs. Interpretable

- Interpretability and Explainability often used interchangeably
- Interpretability is a goal
- Explainability **is one way** of achieving Interpretability

Approaches to explainability

- Build (simple) interpretable models
- Build post-hoc explainer on top of a black-box model
- Influential instances: deletion form the data considerably changes the model's predictions
- Prototypes & criticisms: represent the data distribution

Guidotti et al. A survey Of methods for explaining black box models. ACM Computing Surveys, 2018.

Types of explanations

Global explanations

Explain a models' decision making process in general. Typically: feature importance

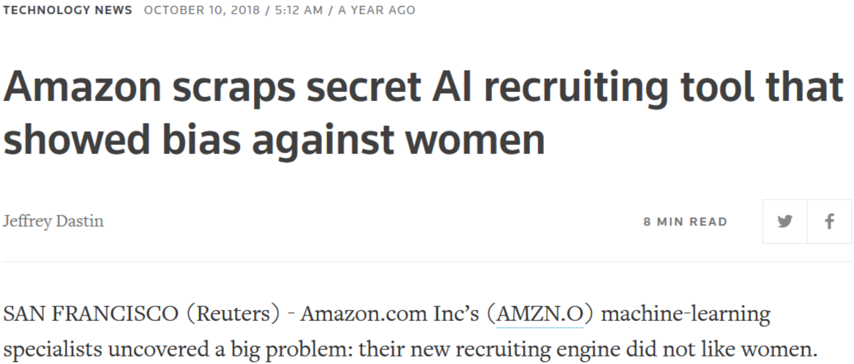
Treeinterpreter, PDP, feature importance

Local explanations

Explain a single prediction. Since it remains challenging to establish fidelity to black box models in globally interpretable approximations, much attention is put on local explanations.

LIME, SHAP, Skater

Explainable AI: why bother?



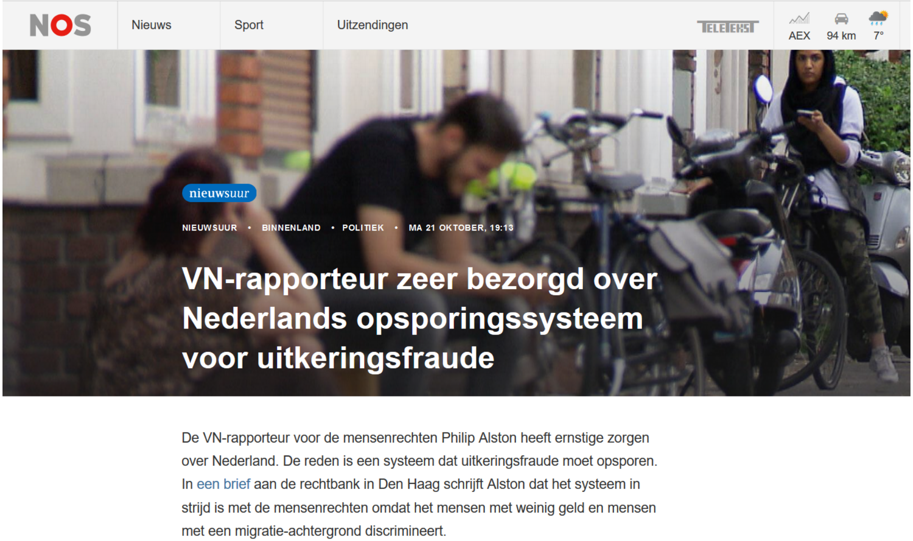
TECHNOLOGY NEWS OCTOBER 10, 2018 / 5:12 AM / A YEAR AGO

Amazon scraps secret AI recruiting tool that showed bias against women

Jeffrey Dastin

8 MIN READ

SAN FRANCISCO (Reuters) - Amazon.com Inc's (AMZN.O) machine-learning specialists uncovered a big problem: their new recruiting engine did not like women.



NIEUWSUUR • BINNENLAND • POLITIEK • MA 21 OKTOBER, 19:13

VN-rapporteur zeer bezorgd over Nederlands opsporingssysteem voor uitkeringsfraude

De VN-rapporteur voor de mensenrechten Philip Alston heeft ernstige zorgen over Nederland. De reden is een systeem dat uitkeringsfraude moet opsporen. In een brief aan de rechtbank in Den Haag schrijft Alston dat het systeem in strijd is met de mensenrechten omdat het mensen met weinig geld en mensen met een migratie-achtergrond discrimineert.

Complex systems raise concern

- Why this ad?
- Why this discount?
- Why this recommendation?

- Why was I rejected?
- Can I change the outcome?
- When will the system fail?

XAI motivators

Model verification

Compliance

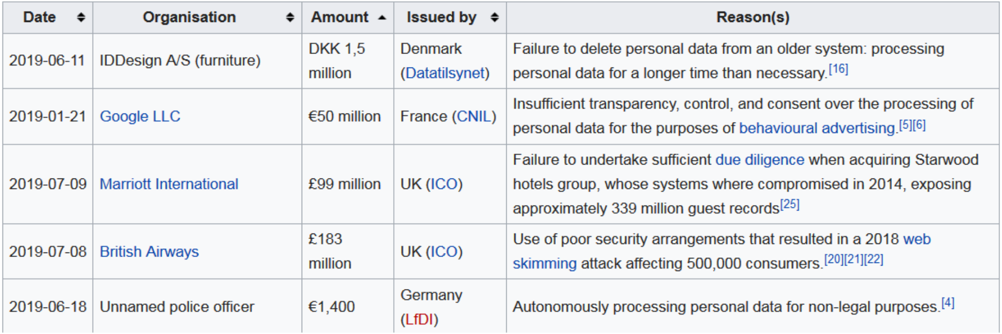
User trust

Actionability

Compliance: GDPR

> “Data subjects have a right to **meaningful information** about the **logic involved** and to the significance and the **envisaged consequence** of automated decision-making”

GDPR fines & notices



Date	Organisation	Amount	Issued by	Reason(s)
2019-06-11	IDDesign A/S (furniture)	DKK 1,5 million	Denmark (Datatilsynet)	Failure to delete personal data from an older system: processing personal data for a longer time than necessary.[16]
2019-01-21	Google LLC	€50 million	France (CNIL)	Insufficient transparency, control, and consent over the processing of personal data for the purposes of behavioural advertising.[5][6]
2019-07-09	Marriott International	£99 million	UK (ICO)	Failure to undertake sufficient due diligence when acquiring Starwood hotels group, whose systems where compromised in 2014, exposing approximately 339 million guest records[25]
2019-07-08	British Airways	£183 million	UK (ICO)	Use of poor security arrangements that resulted in a 2018 web skimming attack affecting 500,000 consumers.[20][21][22]
2019-06-18	Unnamed police officer	€1,400	Germany (LfDI)	Autonomously processing personal data for non-legal purposes.[4]

https://en.wikipedia.org/wiki/GDPR_fines_and_notices

XAI in practice: explaining errors for user trust

Motivational example: predicting next week's sales

- Current model
  - auto-regressors
  - transaction history

- New model(s)
  - ensemble learning
  - 40+ features

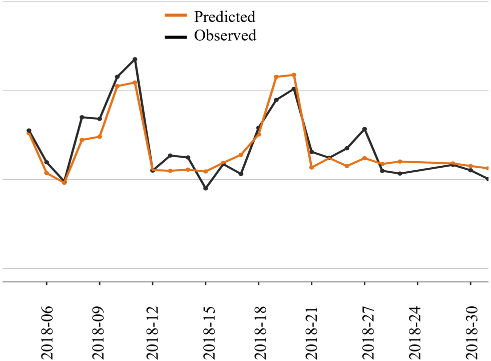
Motivational example: predicting next week's sales

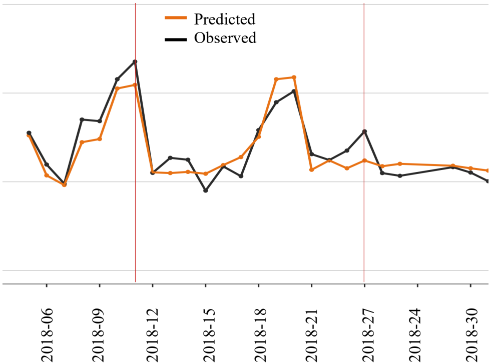
- Current model
  - auto-regressors
  - transaction history

- New model(s)
  - ensemble learning
  - 40+ features

User feedback

- Model perceived as a black-box
- Counter-intuitive results
- Gain in performance vs. loss in interpretability


Predicted
Observed
2018-06
2018-09
2018-12
2018-15
2018-18
2018-21
2018-27
2018-24
2018-30



How can I improve the forecast?

Data Scientist

Can I trust this forecast?

Stakeholder

When does the forecast fail?

End-user

How can we explain the errors of a forecasting model?

A. Lucic, H. Haned, M. de Rijke. Why does my model fail? Contrastive local explanations for retail forecasting. FAT* 2020.

What is a good explanation?

“The key insight is to recognise that one does not explain events per se, but that one explains why the puzzling event occurred in the target cases but not in some counterfactual contrast case.”

“Why A and not B?”

D. J. Hilton. Conversational processes and causal explanation, Psychological Bulletin, 1990.

Explain errors to enhance user trust

- **MC-BRP** Monte Carlo Bounds for Reasonable Predictions
- Identifying unusual properties of a particular observation – we assume large errors occur due to unusual features in the test set that are not present in the training set
- Given an erroneous prediction, MC-BRP generates:
  1. Feature values that would result in a reasonable prediction, based on the *n* most important features
  2. General trends between each feature and the target variable

A. Lucic, H. Haned, M. de Rijke. Why does my model fail? Contrastive local explanations for retail forecasting. FAT* 2020.

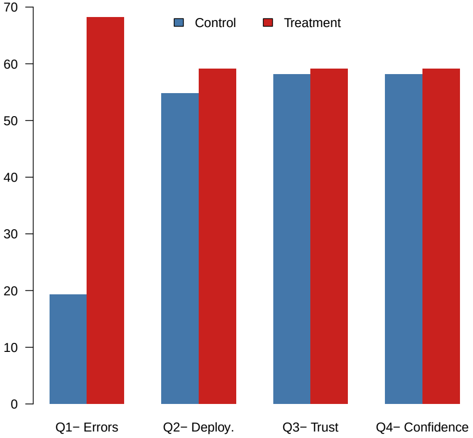
Contrastive explanations for large forecasting errors

Input	Trend	Value	Reasonable range
A	As input increases, sales increase	9628.00	[4140,6565]
B	As input increases, sales increase	18160.67	[8290,15322]
C	As input increases, sales increase	97332.00	[51219,75600]
D	As input increases, sales decrease	226.00	[95,153]
E	As input increases, sales decrease	2013.60	[972,1725]

Contrastive explanations for large errors

We ask our users the following subjective questions:

- **Q1:** I understand why the model makes large errors in predictions
- **Q2:** I would support using this model as a forecasting tool
- **Q3:** I trust this model
- **Q4:** In my opinion this model produces mostly reasonable outputs



Lessons learned

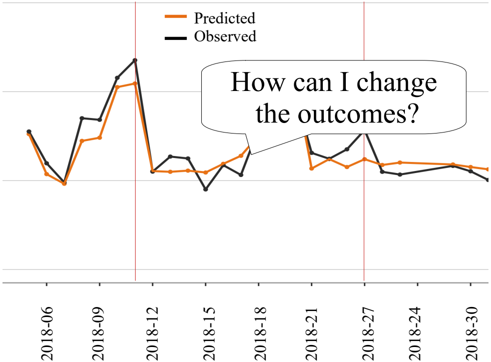
- Explanations generated by our method help users understand why models make large errors
- Explanations do not have a significant impact on support in deploying the model, trust in the model, or perceptions of the model's performance

Algorithmic aversion

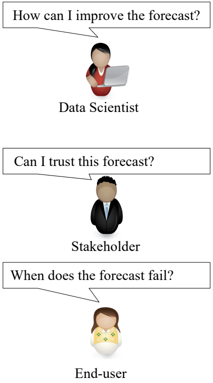
> “We show that people are especially averse to algorithmic forecasters after seeing them perform, even when they see them outperform a human forecaster. This is because people more quickly lose confidence in algorithmic than human forecasters after seeing them make the same mistake”

Dietvorst et al. Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology, 2015.

Explanations are not enough



How can I change the outcomes?



How can I improve the forecast?

Data Scientist

Can I trust this forecast?

Stakeholder

When does the forecast fail?

End-user

Explanations are not enough

> "(...) we found that giving participants the freedom to modify an imperfect algorithm made them feel more satisfied with the forecasting process, more likely to believe that the algorithm was superior, and more likely to choose to use an algorithm to make subsequent forecasts"

Dietvorst et al. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science, 2016.

Actionable explanations: counterfactuals

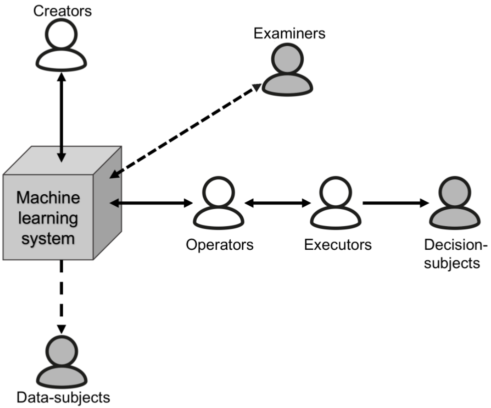
A counterfactual describes the smallest required change to a feature value that changes the prediction to a predefined desired output

- **Model** forecast for next week is 5,000
- **Question** Which feature values must be changed to decrease the forecast to 4,000?
- **Counterfactual** If your delivery on the weekend is no longer free, you will decrease the forecast to below 4,000 transactions

Wachter et al. Counterfactual explanations without opening the black box: Automated decisions and the GDPR, Harvard Journal of Law & Technology, 2018.

XAI as a process rather than a product

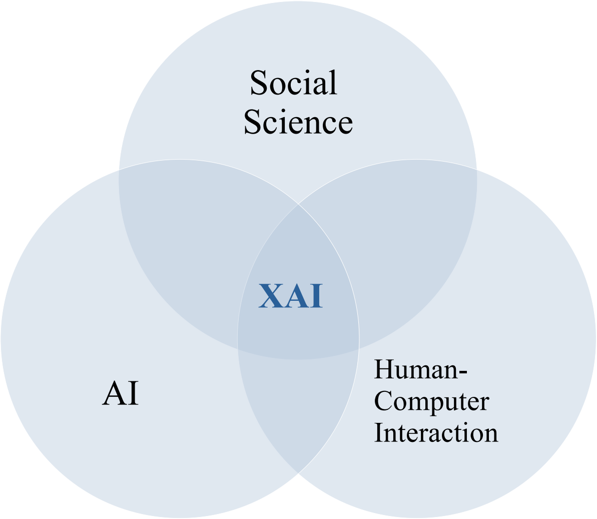
Diverse users



Tomsett et al. Interpretable to Whom? A Role-based model for analyzing interpretable machine learning systems, ICML WHI Workshop, 2018.

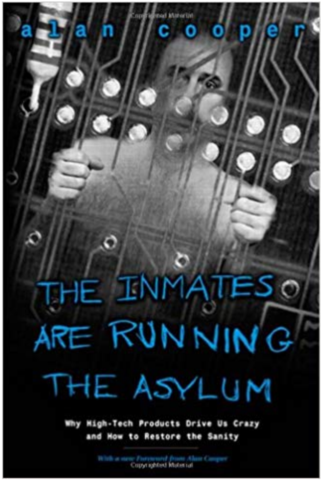
Insights from the social sciences

- Explanations are Contrastive
- Explanations are Selected
- Explanations are Social
- Explanations are Contextual



Users matter

> “Most of us as AI researchers are building explanatory agents for ourselves, rather than for the intended users”


alan cooper
THE INMATES ARE RUNNING THE ASYLUM
Why High-Tech Products Drive Us Crazy and How to Restore the Sanity
With a new Foreword from Alan Cooper

T. Miller et al. Beware of inmates running the Asylum, IJCAI Workshop on explainable AI, 2017.

Thank you

`h.haned@uva.nl`

`github.com/hindantation`