GÖTEBORGS
UNIVERSITET

CHALMERS

GÖTEBORGS
UNIVERSITET

CHALMERS

DAT 550 / DIT 978
Advanced Software Engineering for AI/ML-Enabled Systems

GÖTEBORGS UNIVERSITET
CHALMERS

Lecture 4: On responsible AI/ML Engineering

Your teachers:
Hans-Martin Heyn, Universitetslektor,
Eric Knauss, Docent

Computer Science and Engineering Department, Göteborg University

Become a student representative!

- We need your feedback to enhance and improve this (brand new) course.
- Send a mail to heyn@chalmers.se

What will you learn?

On responsible AI/ML Engineering

• Introduction to Responsible AI/ML	Understand the different aspects of responsible AI/ML Engineering Provide insight into current Legislation for AI in the EU.
• Bias in ML models	Explain why AI/ML models can be biased. Explain different bias avoidance strategies.
• User Interaction and Explainability	Understand the importance of interpretability
• Example on explainability <- Icing of airport runways	Explain the difference between inherently interpretable models and post-hoc explanations

What will you learn?

On responsible AI/ML Engineering

GÖTEBORGS UNIVERSITET

CHALMERS

• Introduction to Responsible AI/ML	Understand the different aspects of responsible AI/ML Engineering Provide insight into current Legislation for AI in the EU.
• Bias in ML models	Explain why AI/ML models can be biased. Explain different bias avoidance strategies.
• User Interaction and Explainability	Understand the importance of interpretability
• Example on explainability <- Icing of airport runways	Explain the difference between inherently interpretable models and post-hoc explanations

What is responsible?

VIII. Conclusion

Google aspires to be a different kind of company. It's impossible to spell out every possible ethical scenario we might face. Instead, we rely on one another's good judgment to uphold a high standard of integrity for ourselves and our company. We expect all Googlers to be guided by both the letter and the spirit of this Code. Sometimes, identifying the right thing to do isn't an easy call. If you aren't sure, don't be afraid to ask questions of your manager, Legal or Ethics & Business Integrity.

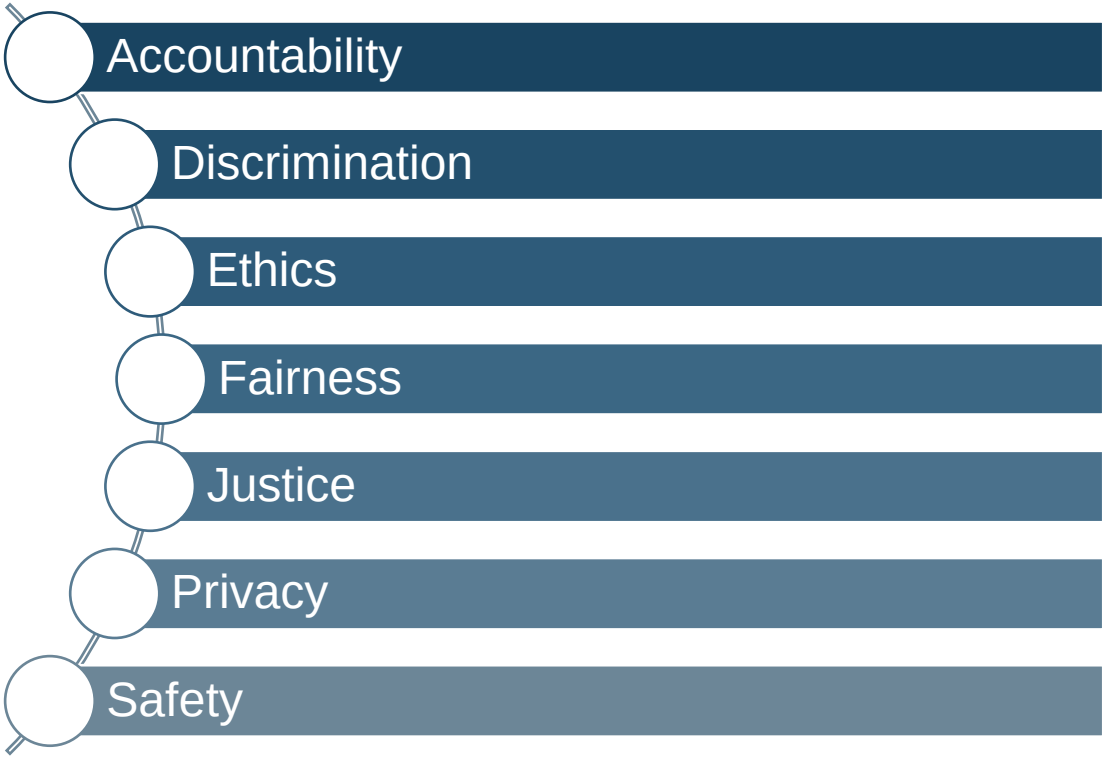
And remember.. don't be evil, and if you see something that you think isn't right – speak up!

Last updated January 24, 2022

Is Google developing responsible products?

Responsibility has many interrelated issues

“It's impossible to spell out every possible ethical scenario”



Each of these issues is a deep and nueanced research topic.

Legal != Ethical

Legal = in accordance with societal laws

- Systematic body of rules governing society; set through government
- Punishment for violation

Ethical = following moral principles of tradition, group, or individual

- Branch of philosophy, science of a standard human conduct
- No legal binding, no enforcement beyond "shame“
- Depends on cultural background
- High ethical standards may yield long term benefits through image

Legal work and ethical behaviour are often entangled
=> Human rights, GDPR and EU Regulation of AI

Challenges

- Misalignment between organizational goals & societal values
  - Financial incentives often dominate other goals ("grow or die")
- Insufficient regulations
  - Poor understanding of socio-technical systems by policy makers
- Engineering challenges, both at system- & ML-level
  - Difficult to clearly define or measure ethical values
  - Difficult to predict possible usage contexts
  - Difficult to prevent malicious actors from abusing the system
  - Difficult to interpret output of ML and make ethical decisions
  - ...

Why do we need to regulate AI?

POLICY AND LEGISLATION | Publication 21 April 2021

Proposal for a Regulation laying down harmonised rules on artificial intelligence

The Commission has proposed the first ever legal framework on AI, which addresses the risks of AI and positions Europe to play a leading role globally.

The Proposal for a Regulation on artificial intelligence was announced by the Commission in April 2021. It aims to address risks of specific uses of AI, categorising them into 4 different levels: unacceptable risk, high risk, limited risk, and minimal risk.

In doing so, the AI Regulation will make sure that Europeans can trust the AI they are using. The Regulation is also key to building an ecosytem of excellence in AI and strengthening the EU's ability to compete globally. It goes hand in hand with the Coordinated Plan on AI.

Downloads

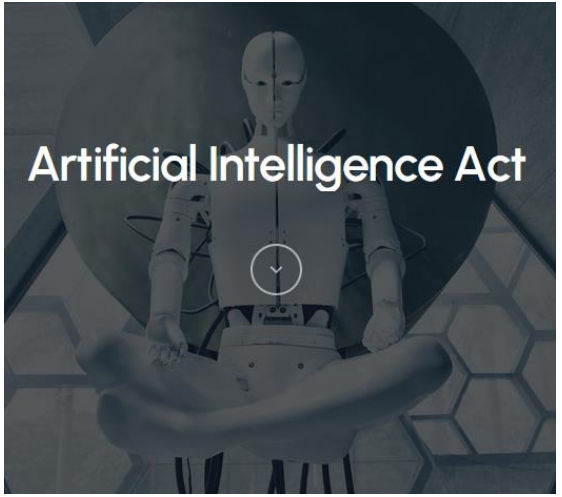
1. Proposal for a Regulation laying down harmonised rules on artificial intelligence (.pdf) Download

See also

View the proposal for a Regulation in all EU languages on EUR-Lex

Related topics

eHealth | Advanced Digital Technologies | Artificial intelligence



Overview EU AI Act

4 classes of AI

https://artificialintelligenceact.eu/

GÖTEBORGS UNIVERSITET

CHALMERS

Prohibited AI Practices (Title II)

- Certain AI Practices (especially involving social scoring) are prohibited.

High-Risk AI Systems (Title III)

- Allowed, but with requirements…
  - Risk management (§9)
  - Data governance (§10)
  - Transparency and Traceability (§11-14)
  - Robustness and Security (§15)
  - Third-party assessment (§19)

Limited Risk AI Systems (Title IV) Transparency & Post-Market Monitoring

Minimal Risk AI Systems

What is an AI? (Title I)

- §3(1) ‘artificial intelligence system’ (AI system) means software that is developed with one or more of the techniques and approaches listed in Annex I and can, **for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with**;

Annex I:

(a) Machine learning approaches, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning;

(b) Logic- and knowledge-based approaches, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems;

(c) Statistical approaches, Bayesian estimation, search and optimization methods.

Prohibited AI Practices (Title II)

§5: Prohibited Artificial Intelligence Practices

1. The following artificial intelligence practices shall be prohibited:
(a) […] AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm;
(b) […] AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm;
(c) […] AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons based on their social behaviour or known or predicted personal or personality characteristics
(d) the use of 'real-time' remote biometric identification systems in publicly accessible spaces […].

High-Risk AI Systems (Title III)

§6 and Annex III

High-risk AI systems pursuant to Article 6(2) are the AI systems listed in any of the following areas:

- Safety critical component (Article 6(1))
- Biometric identification of natural persons
- AI determining access to educational training
- AI intended to be used for recruitment selection
- Access to essential private and public services (e.g. access to loans, emergency service, medical service)
- Law enforcement
- Migration, asylum, border control
- Administration of justice and democratic processes

Transparency Obligations (Title IV)

- Providers shall ensure that AI systems intended to interact with natural persons are designed and developed in such a way that **natural persons are informed that they are interacting with an AI system**, unless this is obvious from the circumstances and the context of use.
  - This obligation shall not apply to AI systems authorised by law to detect, prevent, investigate and prosecute criminal offences, unless those systems are available for the public to report a criminal offence.
- Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful (‘deep fake’), shall disclose that the content has been artificially generated or manipulated.

Harm on society

Harms of allocation:

- Withhold opportunities or resources
- Poor quality of service, degraded user experience

Harms of representation:

- Reinforce stereotypes
- Over/under-representation of certain groups in organisations

	Allocation of resources	Quality of Service	Stereotyping	Denigration	Over- / Under-Representation
Hiring system does not rank women as highly as men for technical jobs	x	x	x		x
Photo management program labels image of black people as “gorillas”		x		x	
Image searches for “CEO” yield only photos of white men on first page			x		x

“The Trouble With Bias”, Kate Crawford, Keynote@N(eur)IPS (2017) and
Challenges of incorporating algorithmic fairness into practice, FAT* Tutorial (2019)

What will you learn?

On responsible AI/ML Engineering

GÖTEBORGS UNIVERSITET

CHALMERS

- Introduction to Responsible AI/ML

Understand the different aspects of responsible AI/ML Engineering

Provide insight into current Legislation for AI in the EU.

- Bias in ML models

Explain why AI/ML models can be biased.
Explain different bias avoidance strategies.

- User Interaction and Explainability
- Example on explainability
  <- Icing of airport runways

Understand the importance of interpretability

Explain the difference between inherently interpretable models and post-hoc explanations

Bias in ML models

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al. Science (2017)

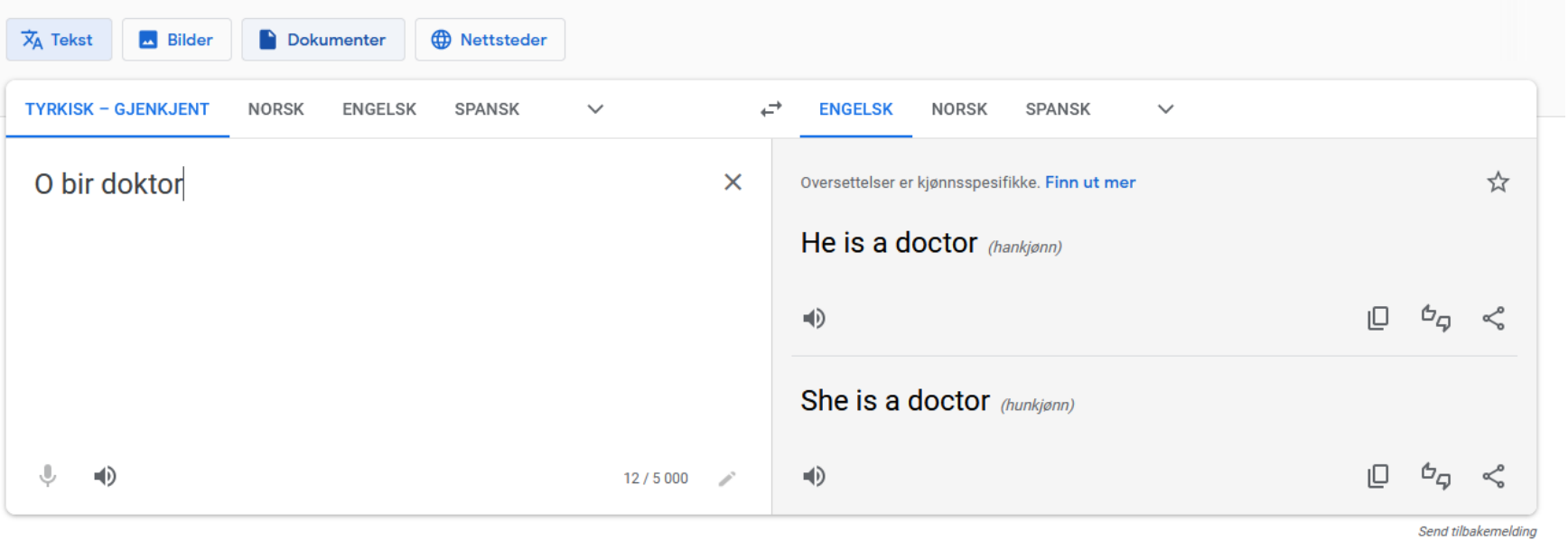
Bias in ML models

Semantics derived automatically from language corpora contain human-like biases, Caliskan et al. Science (2017)

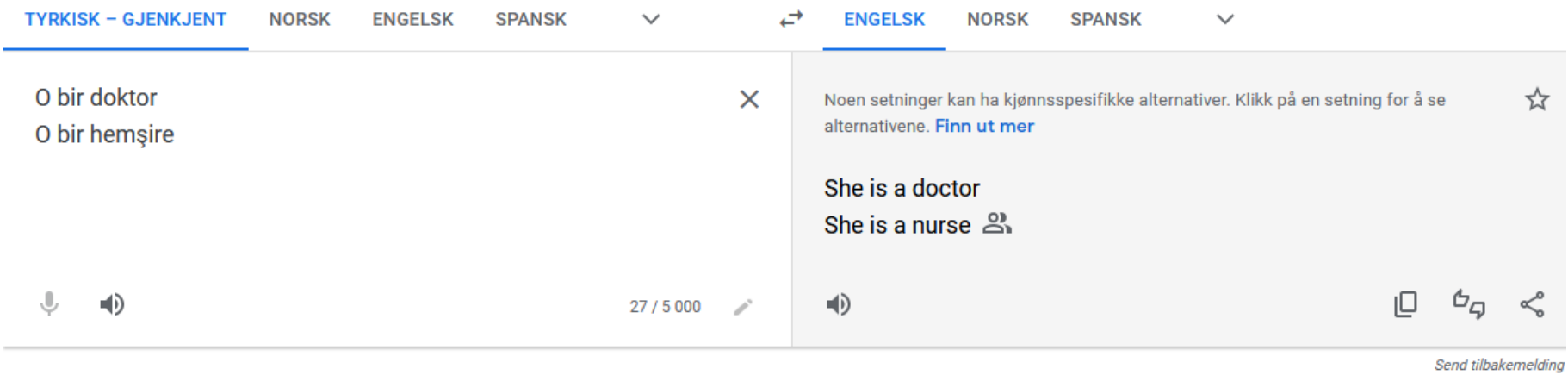
How would you fix that?

https://miro.com/app/board/uXjVMZRmoF4=/?share_link_id=427613621967

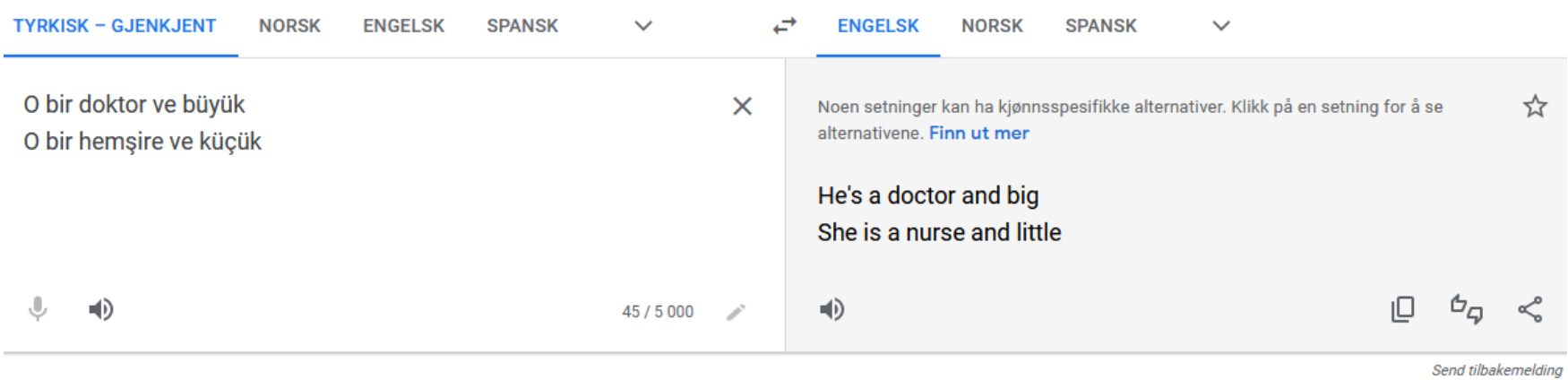
Bias in ML models



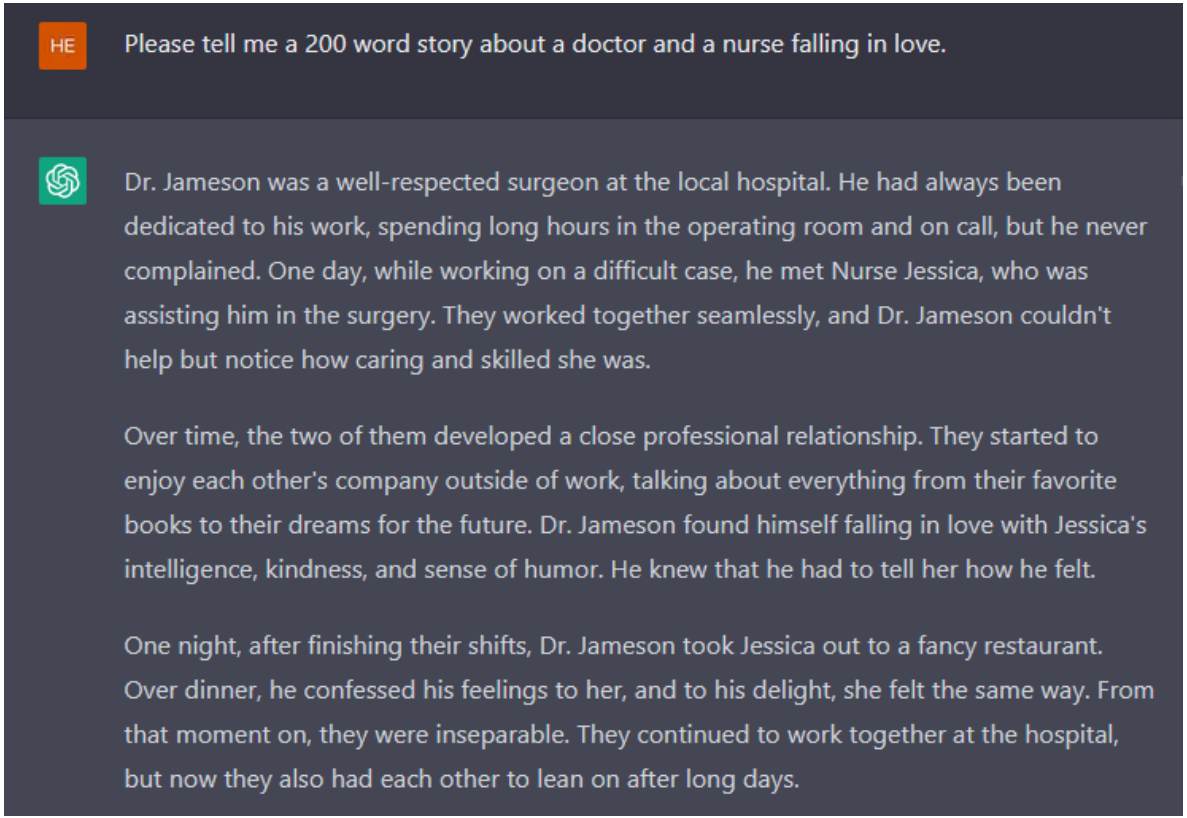
Bias in ML models



Bias in ML models



Bias in ML models



Bias in ML models

HE

Please tell me a 200 word story about a nurse and a doctor falling in love.

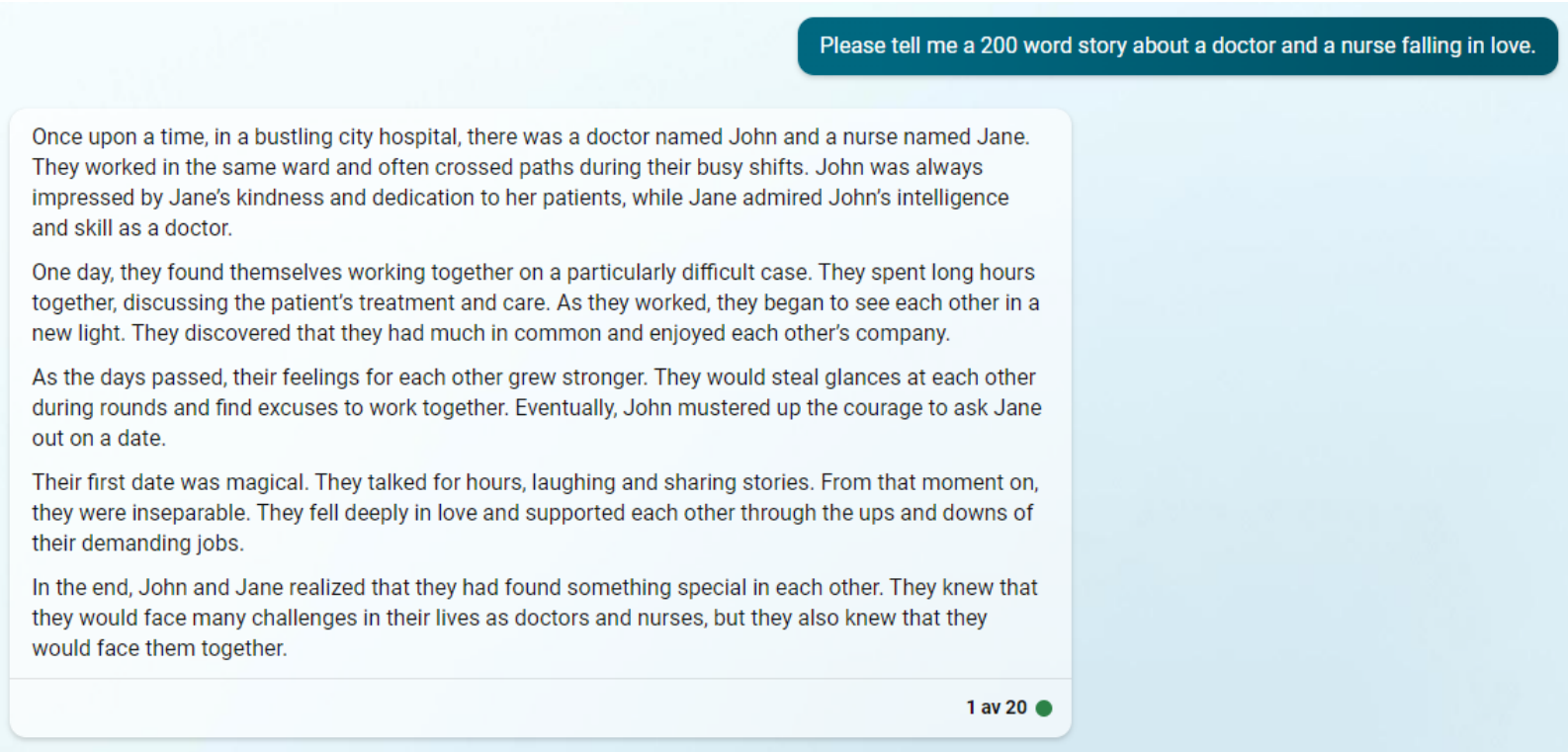
Mary was a nurse who worked tirelessly in the hospital. She had a heart of gold and was always ready to go the extra mile to help her patients. One day, while she was attending to a patient, she met Dr. Mark, a kind and compassionate doctor who had recently joined the hospital. They started working together on several cases, and Mary was impressed by Mark's knowledge and skills.

As they continued to work together, Mary and Mark developed a close friendship. They would often spend their breaks together, discussing their patients and sharing stories about their lives. Mary found herself drawn to Mark's warm smile and his gentle touch, and she realized that she was falling in love with him.

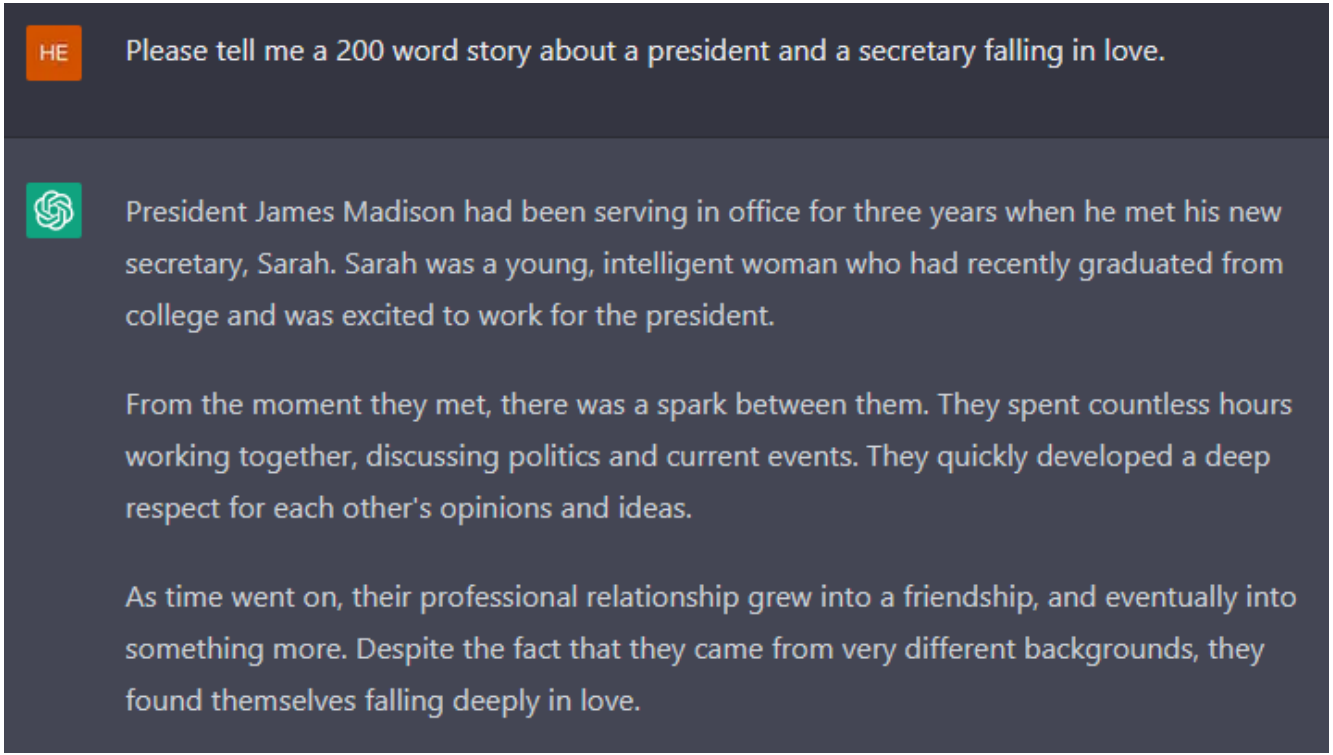
However, Mary was hesitant to confess her feelings to Mark, fearing that it might ruin their friendship and affect their professional relationship. She tried to hide her feelings, but it became increasingly difficult as she spent more time with Mark.

One day, while they were working on a difficult case, Mark confessed his feelings to Mary. He too had been hesitant to express his feelings, but he couldn't keep them bottled up any longer. Mary was overjoyed, and they shared a beautiful moment together.

Bias in ML models



Bias in ML models

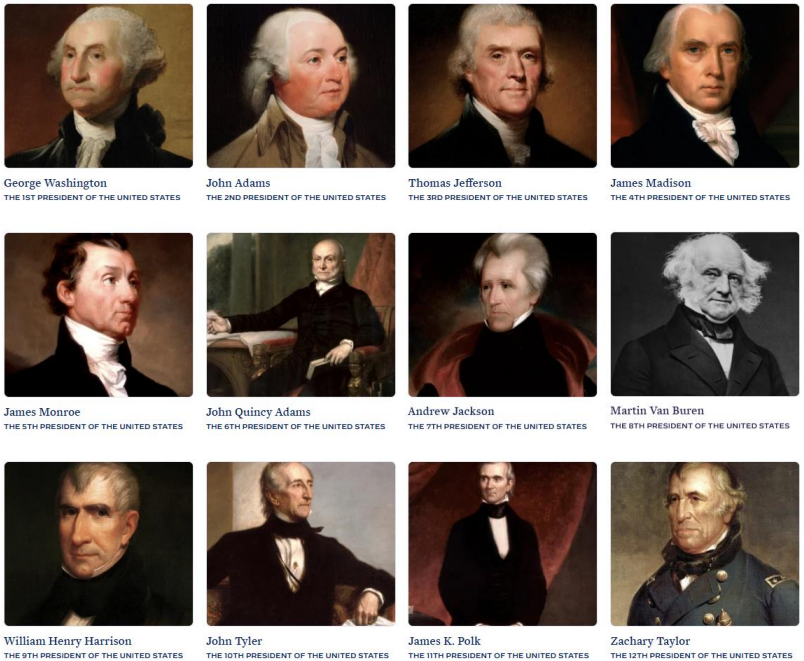


How would you fix that?
Should that be fixed?

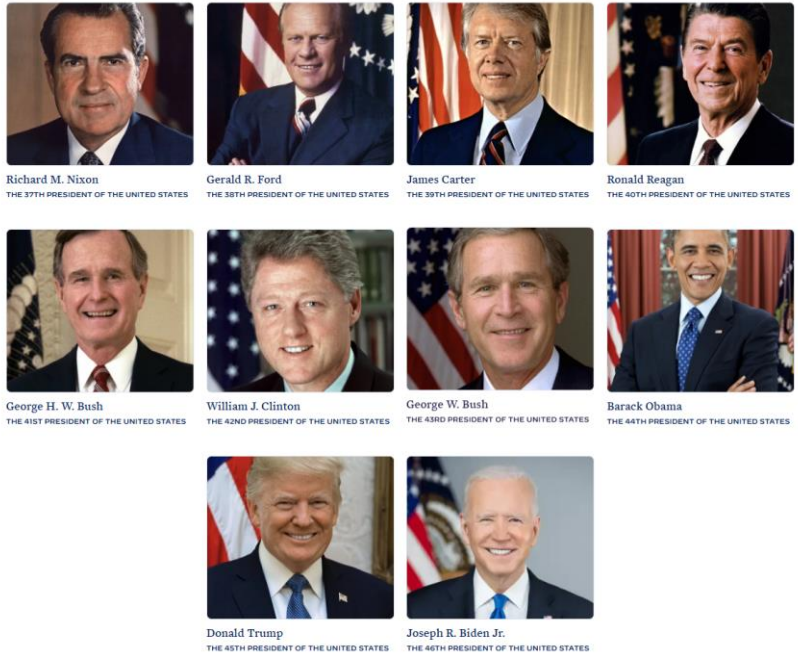
https://miro.com/app/board/uXjVMZRmoF4=/?share_link_id=427613621967

Our world is biased…

…and the training data is sampled from our biased world:



. . .



© https://www.whitehouse.gov/about-the-white-house/presidents/

Our world is biased…

…and the training data is sampled from our biased world:

Where does the bias come from?

- Machine learning learns(!) models from data.
- Bias nearly always stems from the training data, hardly ever from the execution of the ML model itself.

1. Historical bias

- Data reflects the current ”state of the world” which includes a long history of bias.
- For example: How many presidents were women?

2. Biased labelling

- Data labels are often created by humans and reflect personal biases.
- Labels can be indirectly derived from human decisions, such as assigning loan-risk at a bank.

3. Limited features in data

- Some features in data can be predictive for a large part of a population, but not useful or even anti-causal for other parts of the population.

4. Sample size disparity

- Training data is often not equally available for all parts of a target distribution.
- Some parts of the target distribution can be over-represented, allowing the model to better predict outcome for that part of the original population.

5. Proxies (confounder bias)

- Even if we remove sensitive information (e.g, gender, nationality, income) from the data, ML models can find proxy-features that correlate with the removed attributes.

Where does the bias come from?

- Machine learning learns(!) models from data.
- Bias nearly always stems from the training data, hardly ever from the execution of the ML model itself.

What are examples of these biases?
What can we do the prevent them?

1. Historical bias
   - Data reflects the current ”state of the world” which includes a long history of bias.
   - For example: How many presidents were women?
2. Biased labelling
   - Data labels are often created by humans and reflect personal biases.
   - Labels can be indirectly derived from human decisions, such as assigning loan-risk at a bank.
3. Limited features in data
   - Some features in data can be predictive for a large part of a population, but not useful or even anti-causal for other parts of the population.
4. Sample size disparity
   - Training data is often not equally available for all parts of a target distribution.
   - Some parts of the target distribution can be over-represented, allowing the model to better predict outcome for that part of the original population.
5. Proxies (confounder bias)
   - Even if we remove sensitive information (e.g, gender, nationality, income) from the data, ML models can find proxy-features that correlate with the removed attributes.

https://miro.com/app/board/uXjVMZRmoF4=/?share_link_id=427613621967

Time for a break

Get coffee ☺

Not all discrimination is harmful

- Loan risk prediction. Gender discrimination is illegal.
- Medical diagnosis: Gender-specific diagnosis is desirable!

- The problem is **unjustified differentiation** or anti-causal differentiation: Discriminiation on factors that should not matter or that are not causally related.
- Discriminiation is domain-specific and must be analysed in context of the problem domain.

Bias in data

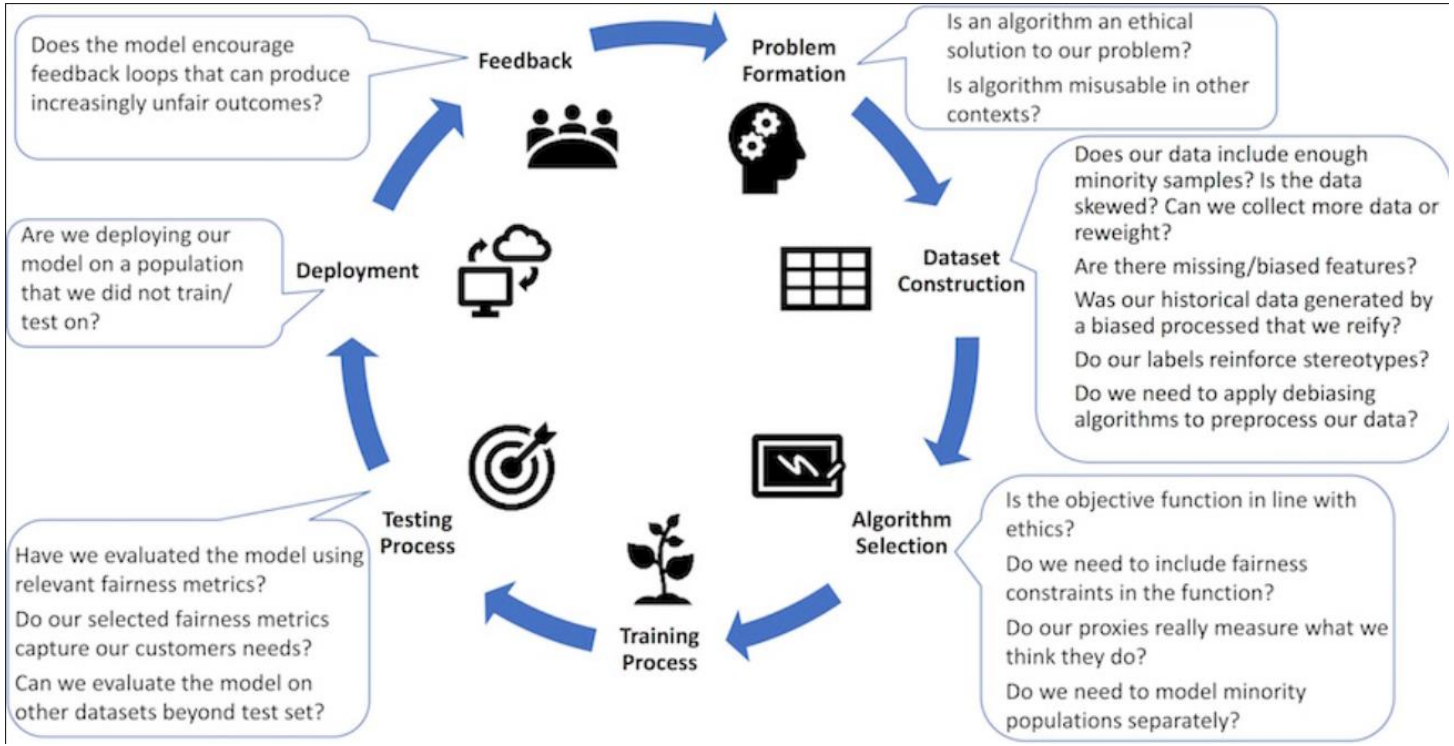
Bias can be introduced at any stage in the data pipeline!

- Machine learning learns(!) models from data.
  - Does the dataset reflect the demographics in the target population?
- Ensure sufficient amount of data for all groups. Avoid over-representation.
- Synthesize data for minority groups to reduce under-represetnation.
  - He is a doctor -> She is a doctor.
- Carefully consider evaluating accuracy accross different groups.

Bennett et al. (2019), Fairness-aware ML https://sites.google.com/view/wsdm19-fairness-tutorial and © Eunsuk Kang (2022)

Bias mitigation in ML model lifecycle

The entire ML lifecycle needs to support Fairness



Bennett et al., Fairness-aware ML https://sites.google.com/view/wsdm19-fairness-tutorial

Bias mitigation through system design

Don’t only focus on the model

- A narrow focus on the model can miss opportunities to increase fairness by
  - redesigning the system around the model,
  - reframing the system goals,
  - mitigating bias with safeguards outside the model.

- Avoid unnecessary distinctions
  - If the distinction between two classes can be biased, consider whether this distinction is actually necessary => Remove gender from data.
- Actively suppress potentially problematic outputs
  - Hard-coded, non-ML (or secondary ML) systems detecting problematic outputs.

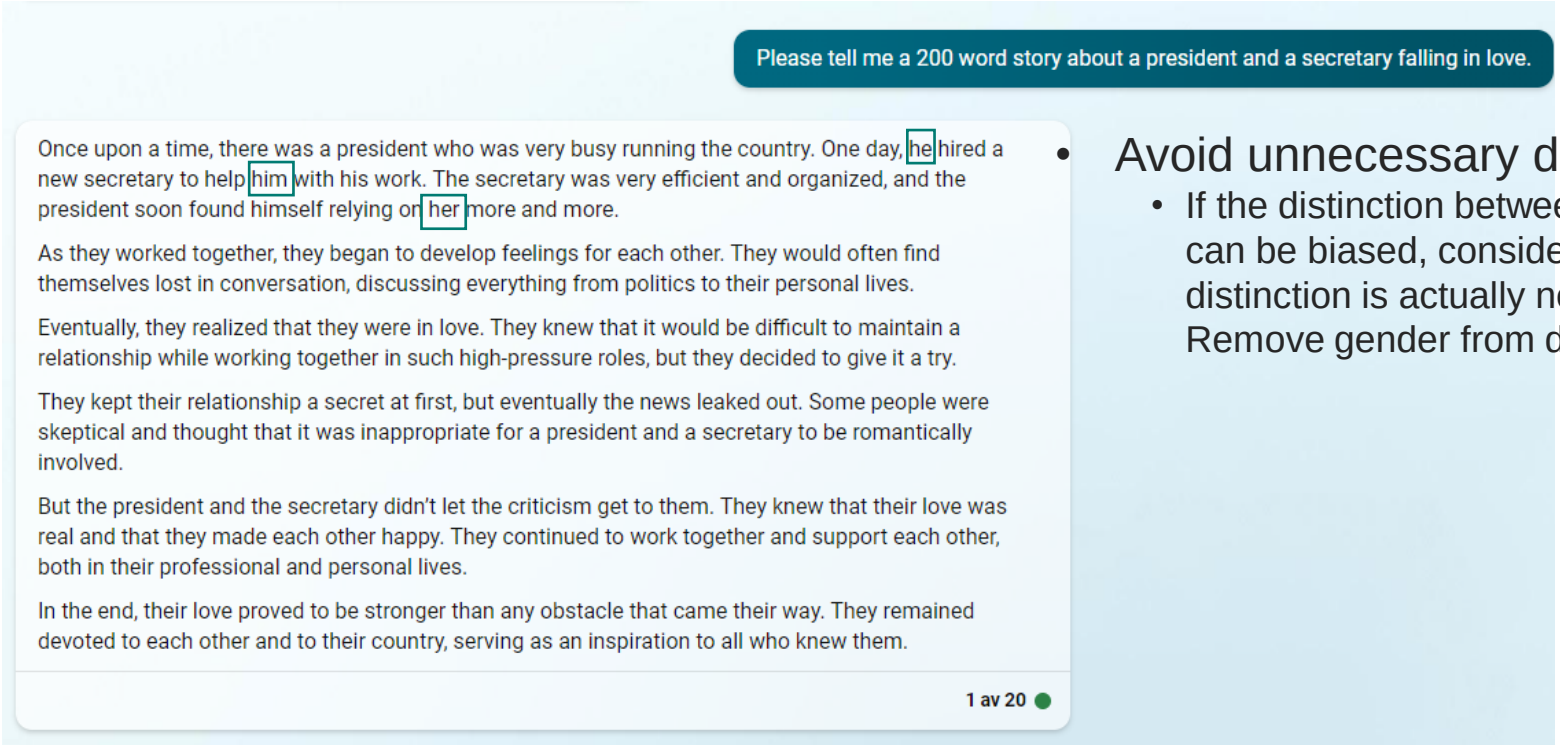
- Fail-soft strategies
  - Minimise the harm caused by discriminatory predictions.

- Keep humans in the loop
  - Involve humans to detect and correct for mistakes and bias.
  - Transparency of outcome.

https://ckaestne.medium.com/fairness-in-machine-learning-and-ml-enabled-products-8ee05ed8ffc4

Bias avoidance

Avoid unnecessary distinctions



- Avoid unnecessary distinctions
  - If the distinction between two classes can be biased, consider whether this distinction is actually necessary => Remove gender from data.

What will you learn?

On responsible AI/ML Engineering

GÖTEBORGS UNIVERSITET

CHALMERS

- Introduction to Responsible AI/ML

Understand the different aspects of responsible AI/ML Engineering

Provide insight into current Legislation for AI in the EU.

- Bias in ML models

Explain why AI/ML models can be biased.
Explain different bias avoidance strategies.

- User Interaction and Explainability

Understand the importance of interpretability

- Example on explainability
  <- Icing of airport runways

Explain the difference between inherently interpretable models and post-hoc explanations

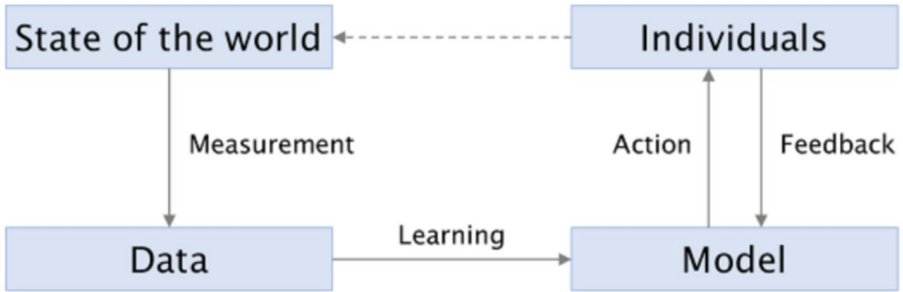
Transparency and Explainability

GÖTEBORGS UNIVERSITET
CHALMERS

High-Risk AI Systems (Title III)

Limited Risk AI Systems (Title IV)

- Keep humans in the loop
  - Involve humans to detect and correct for mistakes and bias.
  - Transparency of outcome.



- ML systems' decisions over time influence the behaviors of populations in the real world.
- But most models are built & optimized assuming that the world is static!
- Difficult to estimate the impact of ML over time.
  - Need to reason about the system dynamics (world vs machine)
  - e.g., what's the effect of a loan lending policy on a population?

Explainability

- Explain how the model made a decision
  - Rules? Cutoffs? Verbal reasoning?
  - What are the relevant factors?
  - Why those rules / cutoffs?
- ML models are complex and based on data
  - Can we understand the rules?
  - Can we understand why rules were found?

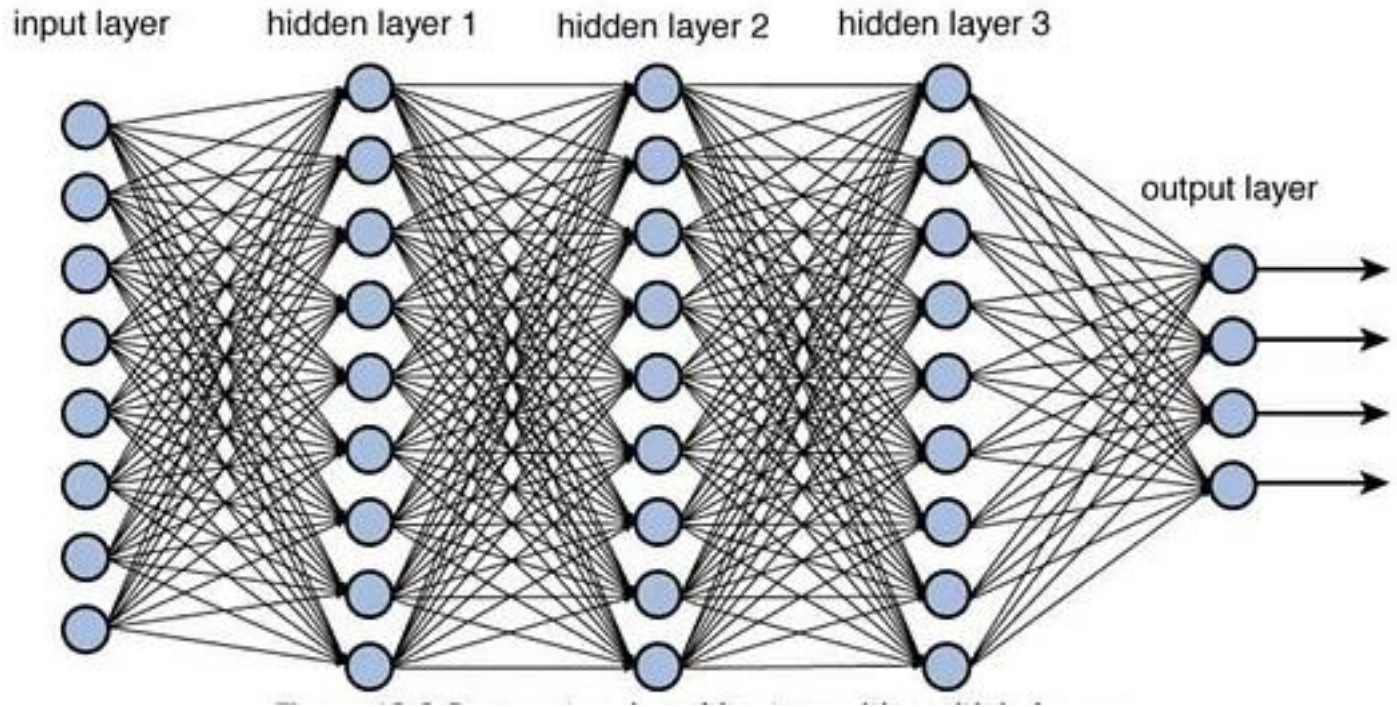
Is this fair?

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE
IF age between 21–23 and 2–3 prior offenses THEN predict arrest
ELSE
IF more than three priors THEN predict arrest
ELSE predict no arrest
```

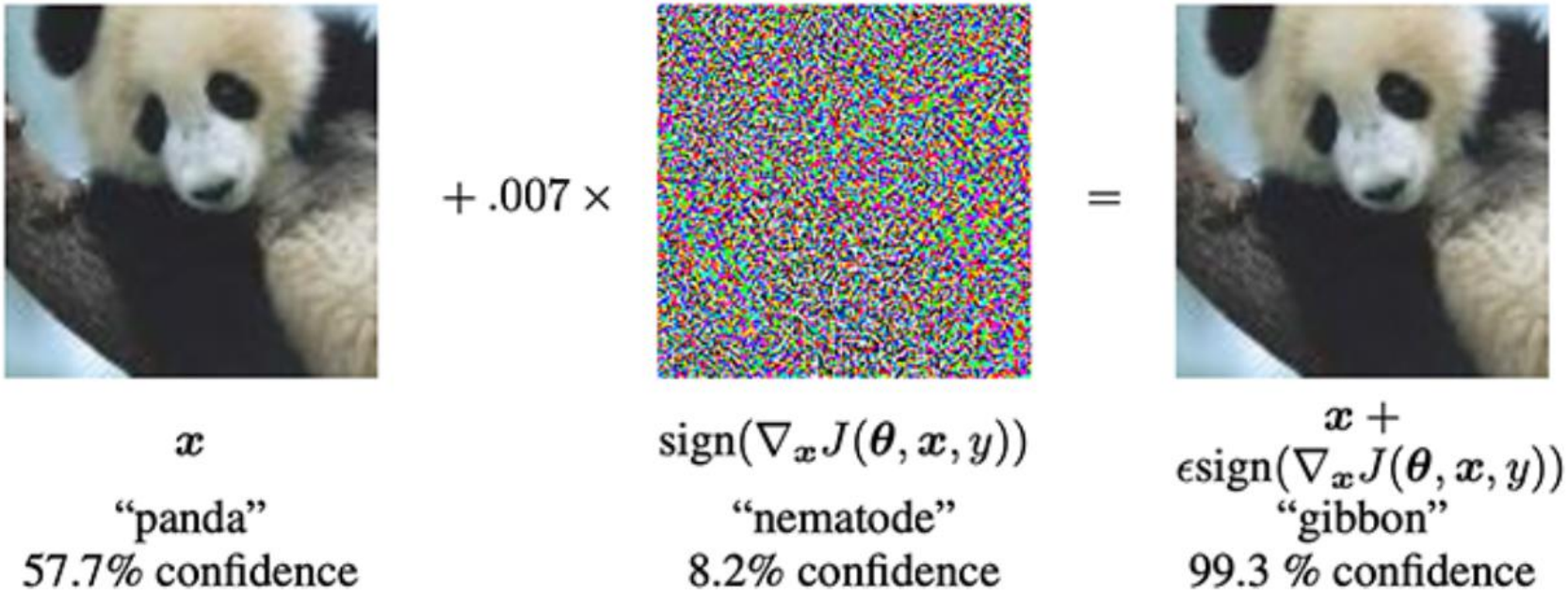
Rudin, Cynthia. "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." Nature Machine Intelligence 1, no. 5 (2019): 206-215.

What's happening here?

And how do we convey the information?

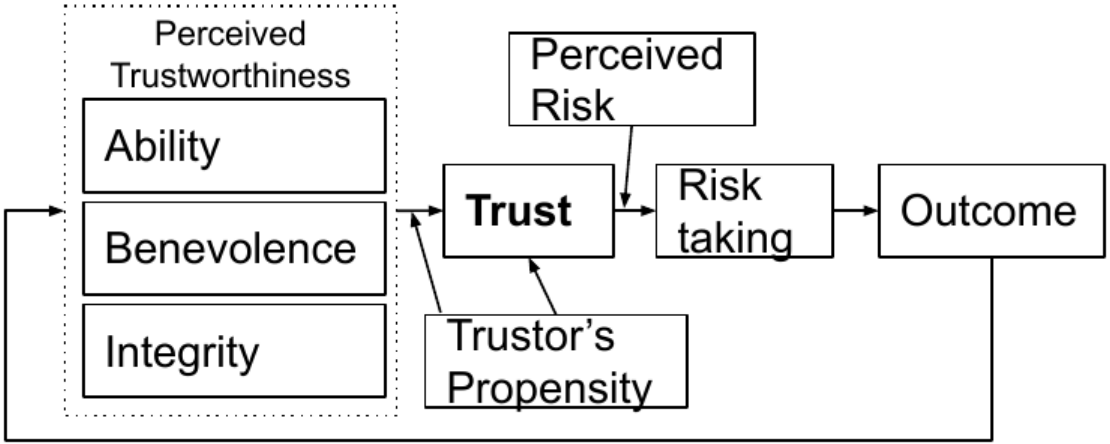


Panda x little Noise = Gibbon



Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.

Establishing trust



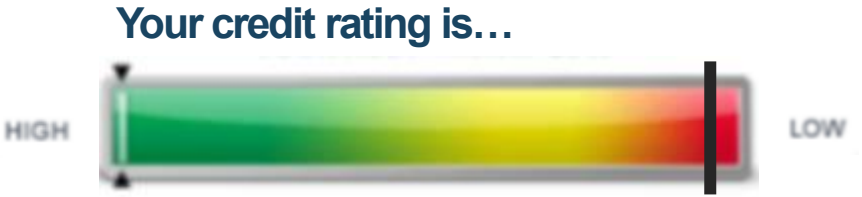
- Willing to accept a prediction more if understandable how it has been made, e.g., model reasoning matches intuition,
- Confidence that the model generalizes beyond target distribution.

Conceptual model of trust: R. C. Mayer, J. H. Davis, and F. D. Schoorman. An integrative model of organizational trust. Academy of Management Review, 20(3):709–734, July 1995.

What is explainability / interpretability?

- Interpretability is the degree to which a human can understand the cause of a decision.
  - We humans love identifying cause-effect relations.
- Interpretability is the degree to which a human can consistently predict the model's results.
  - If we feel we would have made the same decision, we trust the model more.

- (Local) Explainability provides understanding of a single prediction given a certain input.



Your loan application has been declined. If your savings account had had more than 1000 SEK your loan application would be accepted.

- Explainability answers why questions:
  - Why was the loan rejected? (Justification)
  - Why did the treatment not work for the patient? (Debugging)

Interpretability

1.	*Congestive Heart Failure*	1 point		$\cdots$
2.	*Hypertension*	1 point	+	$\cdots$
3.	*Age* $\geq$ *75*	1 point	+	$\cdots$
4.	*Diabetes Mellitus*	1 point	+	$\cdots$
5.	*Prior Stroke or Transient Ischemic Attack*	2 points	+	$\cdots$
	ADD POINTS FROM ROWS 1–5	**SCORE**	=	$\cdots$

SCORE	0	1	2	3	4	5	6
STROKE RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

- In medicine, trust in a diagnosis is established by collecting “indications” that could point towards a medical condition and recommended treatment.
- The idea is that ML models should provide this list of “indications” too.

Rudin, Cynthia, and Berk Ustun. "Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice." Interfaces 48, no. 5 (2018): 449-466.

Intrinsic Interpretability vs. Post-Hoc Explanations?

GÖTEBORGS UNIVERSITET
CHALMERS

- Models can be simplified enough to be understandable / interpretable for humans

```
IF age between 18–20 and sex is male THEN predict arrest
ELSE
IF age between 21–23 and 2–3 prior offenses THEN predict arrest
ELSE
IF more than three priors THEN predict arrest
ELSE predict no arrest
```

1. *Congestive Heart Failure*	1 point		⋯
2. *Hypertension*	1 point	+	⋯
3. *Age* $\geq$ *75*	1 point	+	⋯
4. *Diabetes Mellitus*	1 point	+	⋯
5. *Prior Stroke or Transient Ischemic Attack*	2 points	+	⋯
ADD POINTS FROM ROWS 1–5	**SCORE**	=	⋯

SCORE	0	1	2	3	4	5	6
STROKE RISK	1.9%	2.8%	4.0%	5.9%	8.5%	12.5%	18.2%

- Explanations are provided for ”black-box model” decisions; locally or globally

Your loan application has been declined.

If your savings account had had more than 1000 SEK your loan application would be accepted.

Loan applications are always declined if the saving account has less than 1000 SEK.

What would you prefer? Intrinsic interpretability or post-hoc explanations?

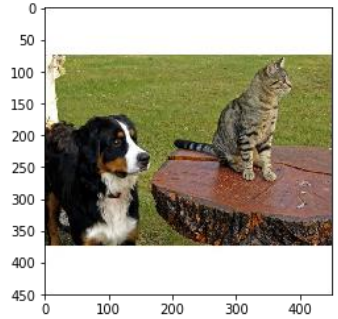
Not all linear models and decision trees are inherently interpretable!

- Models can be very big, many parameters.
- Nonlinear interaction is hard to model in linear models.
- Random forest models are no longer understandable, because they average over multiple different tree instances.

```
173554,681081086 * root + 318523,818532818 * heuristicUnit +
-103411,870761673 * eq + -24600,5000000002 * heuristicVsids +
-11816,7857142856 * heuristicVmtf + -33557,8961038976 *
heuristic + -95375,3513513509 * heuristicUnit * satPreproYes +
3990,79729729646 * transExt * satPreproYes + -136928,416666666
* eq * heuristicUnit + 12309,4990990994 * eq * satPreproYes +
33925,0833333346 * eq * heuristic + -643,428571428088 *
backprop * heuristicVsids + -11876,2857142853 * backprop *
heuristicUnit + 1620,24242424222 * eq * backprop +
-7205,2500000002 * eq * heuristicBerkmin + -2 * Num1 * Num2 +
10 * Num3 * Num4
```

Post-Hoc explanation: Feature importance

- The idea is to permute (remove) a feature value in training and validation datasets to not use it for prediction.
  - Then measure influence on accuracy of output.
- Effect for feature + interaction.
- However, features can be correlated.
  - Removing one feature might not remove the effect of that feature on the output!! => Causality
- Can be crazy expensive because you need to train and validate the model for each permutation.



Computer says: It's a cat!

Source: https://github.com/marcotcr/lime

Time for a break

Get more coffee ☺

What will you learn?

On responsible AI/ML Engineering

GÖTEBORGS UNIVERSITET

CHALMERS

- Introduction to Responsible AI/ML

Understand the different aspects of responsible AI/ML Engineering

Provide insight into current Legislation for AI in the EU.

- Bias in ML models

Explain why AI/ML models can be biased.
Explain different bias avoidance strategies.

- User Interaction and Explainability
- Example on explainability
  <- Icing of airport runways

Understand the importance of interpretability

Explain the difference between inherently interpretable models and post-hoc explanations

Example Explainability:

Predicting runway conditions using XGBoost and explainable AI

GÖTEBORGS UNIVERSITET
CHALMERS

- Snow and ice on airport runways are a problem (especially here in the Nordic countries).
- Reduction of available tire-pavement friction.
    - Safety threat for aviation.
- Economic costs associated with de-icing.
- Runway closure needs to be carefully decided. Huge impact on travelers / airport.

© Esseco UK DE-Icing

“Pilots [and ground-crew] need accurate and timely information on actual runway surface conditions to active appropriate safety procedures.”

Slippery runways on airports…

Longyearbyen, Svalbard, 2015

Problem definition:

Predicting runway conditions using XGBoost and explainable AI

GÖTEBORGS UNIVERSITET
CHALMERS

$$\mu_B = \frac{D_{\text{brakes}}}{mgcos(\varepsilon) - L}$$

m: Mass of aircraft

μ_B: Brake coefficient

D_{brakes}: Brake force

L: Aerodynamic lift

ε: Slope of runway

© Honeywell

- The pilots need to set the desired D_{brakes} of the aircraft before landing according to the expected friction on the runway.
  - And that friction can be limited due to ice, snow and slush.
  - Therefore, the brake force needs to be limited too.

Motivation:

Predicting runway conditions using XGBoost and explainable AI

Cold Regions Science and Technology 199 (2022) 103556

Contents lists available at ScienceDirect

Cold Regions Science and Technology

journal homepage: www.elsevier.com/locate/coldregions

A decision support system for safer airplane landings: Predicting runway conditions using XGBoost and explainable AI

Alise Danielle Midtfjord *, Riccardo De Bin, Arne Bang Huseby

University of Oslo, Department of Mathematics, 0851 Oslo, Norway

- The idea here is to use a machine learning model to “create a combined runway assessment system”.
- The system includes a classification model to identify slippery conditions.
- The system also includes a regression model to predict the level of “slipperiness”.
- The output is presented to the airport tower crew who make the ultimate decision (inform pilot, initiate de-icing, close runway). => Decision support system

Data Sources:

Predicting runway conditions using XGBoost and explainable AI

GÖTEBORGS UNIVERSITET
CHALMERS

- Weather data from measurement devices at the airport
    - Wind speed, Temperature, Humidity, Precipitation
- Manually written runway reports (Snowtam) (2-3 times / day)
    - Created by airport operator. Includes descriptive information about runway contamination.
- Historic Flight data
    - Includes data from Boeing 737-600/700/800 landing at Oslo Gardamoen (SAS & Norwegian)
    - Data includes acceleration, brake pressure, flap position, engine thrust

Table 2
Number of landings at Oslo Airport in our dataset for the winter seasons 2009/2010 until 2018/2019.

Class	Description	Number of landings
Non-slippery	Non-friction limited	193,056
	Friction limited and $\mu_B > 0.15$	2,289
Slippery	Friction limited and $\mu_B \leq 0.15$	5,163

Based on your intuition, do you foresee any problems with the datasets?

Data preparation:

Predicting runway conditions using XGBoost and explainable AI

GÖTEBORGS UNIVERSITET
CHALMERS

- Weather data need to capture “weather trend”:
  - Include time-lag (kind of sliding window, with discrete steps (1,3,6,12,24) hours back.
- Code the Snowtam reports. Codes describe the runway conditions and contaminiation:
- All together, the dataset contains 151 (!!) variables, including time lags, trends and one-hot-encoding of Snowtam reports

Table 3
Contamination codes and types reported in the Snowtam reports.

Code	Description
0	Bare and Dry
1	Damp
2	Wet
3	Rime
4	Dry Snow
5	Wet Snow
6	Slush
7	Ice
8	Compacted or rolled snow
9	Frozen ruts or ridges

Model training: XGBoost

Predicting runway conditions using XGBoost and explainable AI

GÖTEBORGS UNIVERSITET
CHALMERS

- XGBoost: eEXtreme Gradient Boosting is a scalable implementation of gradient boosting decision trees.
  - Supervised ML method based on decision trees and iteratively minimizing a loss function.

$$\text{obj}(f_k(\boldsymbol{x})) = \sum_{i=1}^{n} L\left(y_i, \widehat{f}(\boldsymbol{x}_i)^{[k-1]} + f_k(\boldsymbol{x}_i)\right) + \Omega(f_k(\boldsymbol{x}))$$

Avoids too complex decision trees

$$\Omega(f_k(\boldsymbol{x})) = \gamma T_k + \frac{1}{2}\lambda \|w_k\|^2$$

Number of ”tree leafes”

Magnitude of weights

Model training: XGBoost

Predicting runway conditions using XGBoost and explainable AI

GÖTEBORGS UNIVERSITET
CHALMERS

- XGBoost: eEXtreme Gradient Boosting is a scalable implementation of gradient boosting decision trees.
  - Supervised ML method based on decision trees and iteratively minimizing a loss function.

$$\text{obj}(f_k(\boldsymbol{x})) = \sum_{i=1}^{n} L\left(y_i, \hat{f}(\boldsymbol{x}_i)^{[k-1]} + f_k(\boldsymbol{x}_i)\right) + \Omega(f_k(\boldsymbol{x}))$$

Avoids too complex decision trees

$$\Omega(f_k(\boldsymbol{x})) = \gamma T_k + \frac{1}{2}\lambda \|w_k\|^2$$

Number of ”tree leafes”

Magnitude of weights

How to “tune” the parameters of the penalty term?

Parameter tuning

Magic happens.

- They used a 10-fold nested cross validation.
- For each fold, 20 random combination of parameters were sampled from a distribution (see below), trained with 2/3 of the fold's data and tested with the remaining 1/3. The best in each fold was selected.

Table 4
Model parameters that where tuned together with the distributions they were sampled from.

Parameter	Explanation	Distribution
n_estimators	Number of trees	{50,250}
reg_lambda	λ	$U(0,10)$
min_split_loss	γ	$U(0,0.4)$
subsample	Subsample ratio	$U(0.3,1)$
learning_rate	Step size shrinkage	$U(0.1,0.21)$

Results

- Runway & Scenario: Physics-based prediction models (much more limited)
- Snowtam: Human assessment of airport ground-crew

Table 5
Confusion matrices for the prediction from the different methods, where the highest number of TP and TN is marked in green and the lowest in red.

		XGBoost		Runway		Scenario		Snowtam	
		Slippery	Non-Slippery	Slippery	Non-Slippery	Slippery	Non-Slippery	Slippery	Non-Slippery
Actual	Slippery	**4 740**	423	**3 905**	1 258	4 223	940	4 006	1 157
	Non-Slippery	28 863	166 482	46 967	148 378	78 894	**116 451**	20 679	**174 666**

Metric	XGBoost	Runway	Scenario	Snowtam
Sensitivity	**0.918**	**0.756**	0.818	0.776
Specificity	0.852	0.760	**0.596**	**0.894**
G-Mean	**0.885**	0.758	**0.698**	0.833

Can we trust the model?

Unlike physics-based models, there is no explanation of the assessment

- Although not a deep neural network, the created model is complex. It combines scores from 50-250 decision trees.
    - It is difficult to impossible to trace back all of the trees.
- SHAP (Shapley Additive Explanations) is a method to create local (!) explanations for ML models.
    - Local: Why a specific (local) observation got its prediction value.
    - It uses game theory (long story, see paper below) to provide these explanation

Lundberg, S., Erion, C., H, G., et al., 2020. From local explanations to global understanding with explainable ai for trees. Nat. Mach. Intell. 2, 56–67.
Shapley, L.S., 2016. 17. A Value for N-Person Games. Princeton University Press, pp. 307–318.

Shapley Additive Explanations

In a nutshell

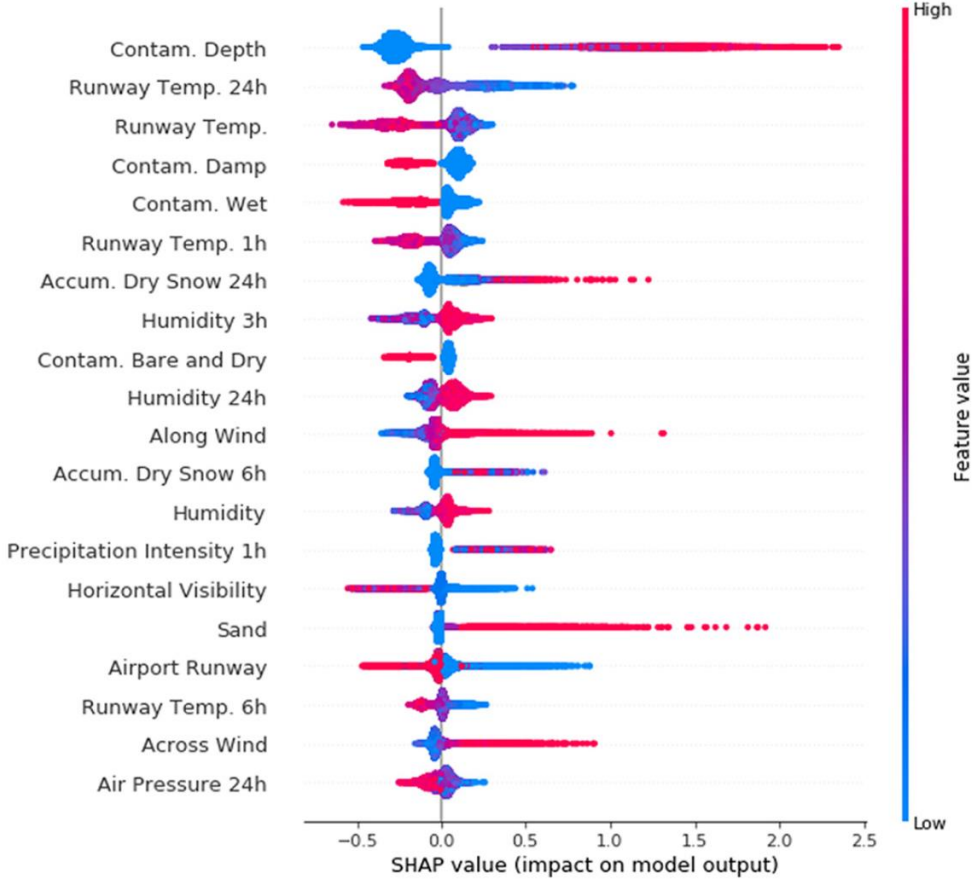
- All input variables are “players in a game”.
- The game is to predict the state of the runway using the available players.
- In the game, individual players can be removed ($z_j \in \{0,1\}$).
- All players together contribute to the final “score”, i.e., the prediction output.

$$g(\mathbf{z}) = \phi_0 + \sum_{j=1}^{m} \phi_j z_j,$$

- The problem is that variables can be correlated. This can lead to ”spurious association”, i.e., a variable explains something that it shouldn’t.
- To break these dependencies, they used rules of causal inference.

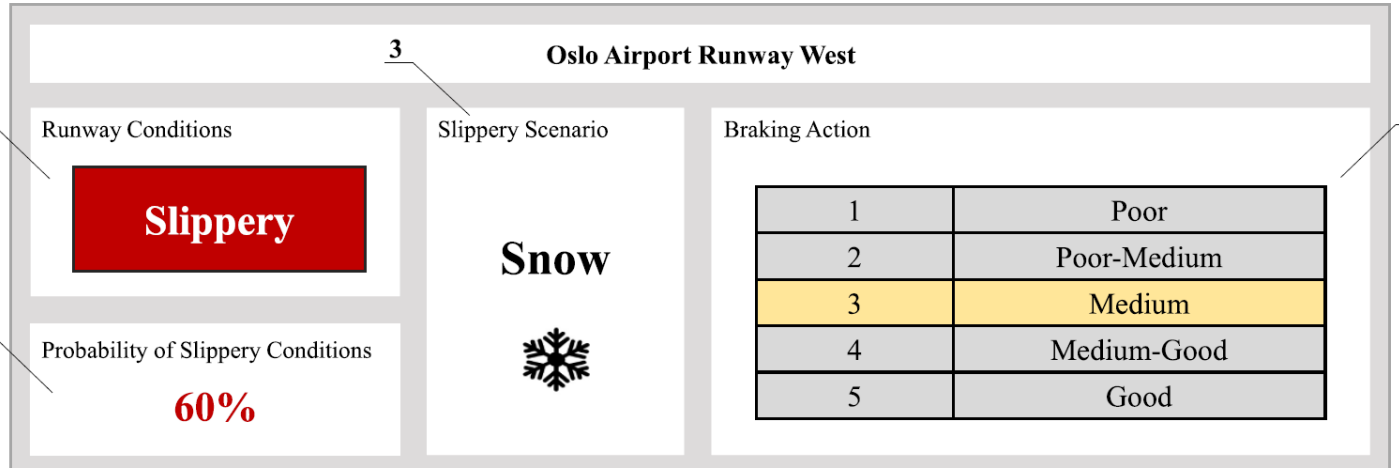
Lundberg, S., Erion, C., H, G., et al., 2020. From local explanations to global understanding with explainable ai for trees. Nat. Mach. Intell. 2, 56–67.
Pearl J., 2019, Causal Inference in Statistics: A Primer’
Shapley, L.S., 2016. 17. A Value for N-Person Games. Princeton University Press, pp. 307–318.

Shapley Additive Explanations



- SHAP values provide information about why a single prediction happened.
  - But applying it over the entire test-dataset can give an idea about global explanations.
- An increase in SHAP value (moving right on x-axis) contributes to a higher probability of “slippery”.
  - A decrease contributes to lowering the probability of “slippery”.
- 20 most influential variables (out of 151).

GUI for tower crew



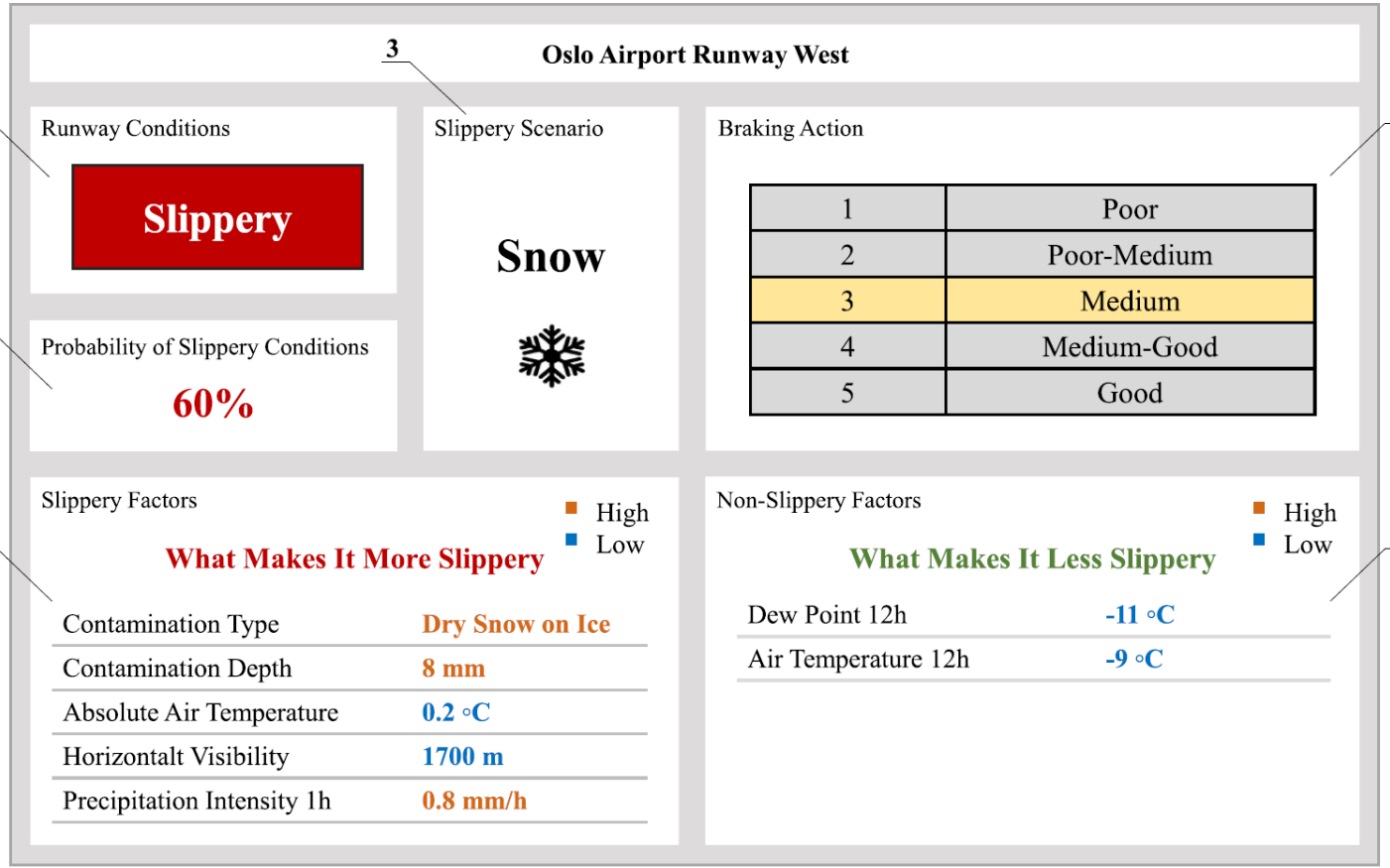
Would you as tower operator trust this information?

1,2: Output classification model

3: Scenario model

4: Output regression model (not discussed)

GUI for tower crew



1,2: Output classification model

3: Scenario model

4: Output regression model (not discussed)

5,6: SHAP value output

What will you learn?

On responsible AI/ML Engineering

• Introduction to Responsible AI/ML	Understand the different aspects of responsible AI/ML Engineering Provide insight into current Legislation for AI in the EU.
• Bias in ML models	Explain why AI/ML models can be biased. Explain different bias avoidance strategies.
• User Interaction and Explainability	Understand the importance of interpretability
• Example on explainability <- Icing of airport runways	Explain the difference between inherently interpretable models and post-hoc explanations

GÖTEBORGS
UNIVERSITET

CHALMERS