Freie Universität Berlin

AI in the wild

Prof. Dr. Eirini Ntoutsi

Ringvorlesung "Mathematikerinnen und Informatikerinnen heute", Berlin, 24/05/2022

To know us better ☺

- 09/1996 – 09/2001, Diploma, Computer Engineering and Informatics (CEID), Polytechnic School, University of Patras, Greece
  - ~150 students (~30 females)
  - Almost only male professors
  - Stereotypes of CS students “stay up late coding and have no social life”
  - Student assistant job at CTI: C4.5 (decision tree algorithm) modification
  - Diploma thesis: AI Games: “*Modeling and improvement of human ability in strategy games*”, primary supervisor: Dimitris Kalles
  - First (and all time favourite) book I read

- Software engineer in a startup during the last year of my studies

D. Kalles, A. Papagelis, E. Ntoutsi, "*Induction of decision trees in numeric domains using set-valued attributes*", Intelligent Data Analysis (IDA), 3(4), 323-347, 2000.

D. Kalles and E. Ntoutsi, "*Interactive Verification of Game Design and Playing Strategies*", 14th IEEE International Conference on Tools with Artificial Intelligence, Washington DC,2002.

To know us better ☺

- 09/2001 – 09/2003 MSc, Computer Science, Polytechnic School, University of Patras, Greece
  - Master thesis: Text Mining: "*Data mining from news data and association with real data*", supervisor: Dimitris Christodoulakis
  - In parallel, working as a full-time software engineer at CTI
- 09/2003 – 09/2008 PhD in Data Mining, University of Piraeus, Athens, Greece
  - PhD topic: "*Similarity Issues in Data Mining - Methodologies and Techniques*", supervisor: Yannis Theodoridis
  - PhD scholarship from HERACLETOS EPEAEK II Programme (2003-2005) supported by the Greek Ministry of Education and the EU
  - PhD researcher for the project GeoPKDD (Geographic Privacy-aware Knowledge Discovery and Delivery) (FP6/IST, 2005-09)
- 04/2007 – 02/2009 Co-Founder and AI expert, NeeMo Startup, Greece
  - After winning the 1st prize in National Innovation 2006 Competition

To know us better ☺

- 09/2008 – 12/2008, short visit at LMU Munich, Group of Prof. Kriegel
  - Alexander von Humboldt postdoc fellow application
  - Accepted but didn't start it immediately
- 2009 Data Mining Expert, National Hellenic Organization (OTE), Athens
  - (first time) female-dominated team, a totally different experience
- 02/2010 - 01/2012, Alexander von Humboldt postdoc fellow, Institute for Informatics, LMU Munich, Germany
  - Topic: High dimensional streams
- 02/2012 - 02/2016, post-doctoral researcher & lecturer, Institute for Informatics, LMU Munich, Germany
  - Project Transalpine mobility and cultural transfer, Research Unit of the German Research Foundation (FOR 1670)
- Female PhD/Postdoc mentoring program, Frauenbeauftragte, LMU

Alexander von Humboldt
Stiftung/Foundation

ARCHAEOBIOCENTER

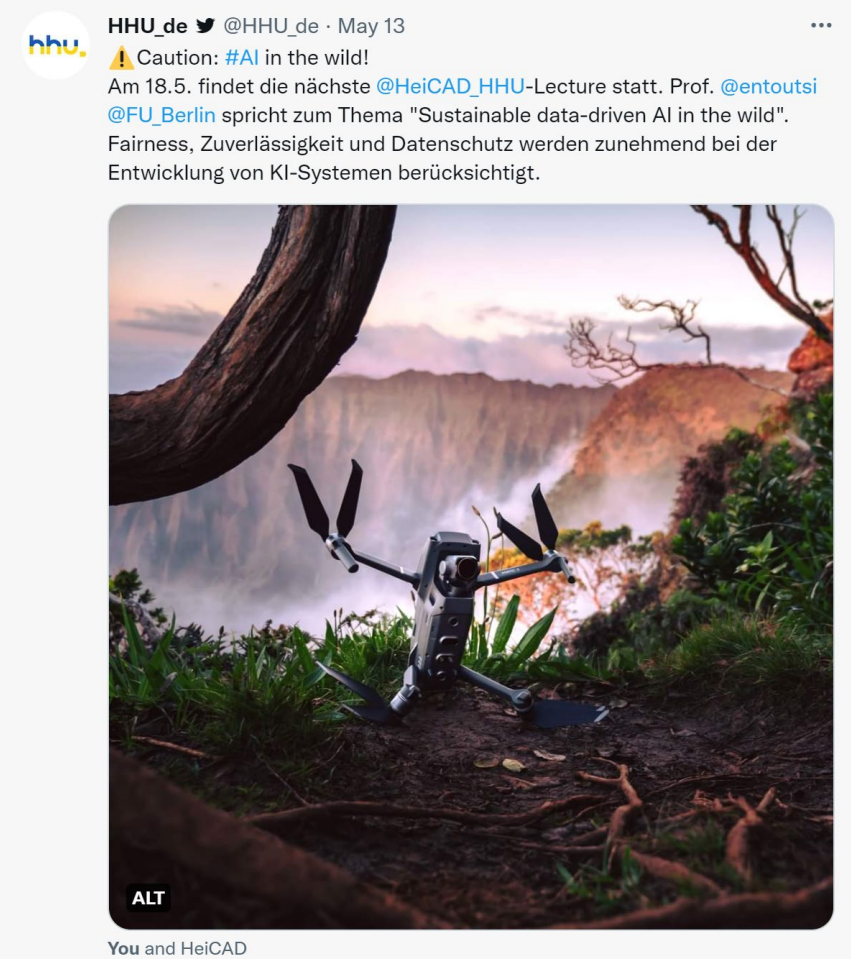
To know us better ☺

- 03/2016 – 02/2021, Associate Professor, Faculty of Electrical Engineering & Computer Science Leibniz University Hannover
  - Member of the L3S Research Center (since May 2016)
  - Female professors mentoring program, Frauenbeauftragte, LUH
- Since March 2021, Full Professor, Free University of Berlin
  - During the pandemic, got involved in several gender-related events
    - 1st Greek ACM-W Chapter Winter School on Fairness in AI

“AI in the wild” - abstract

Algorithmic-based decision-making powered via AI and (big) data has already penetrated into almost all spheres of human life, from recommendations and healthcare to predictive policing, university admission, and autonomous driving, allowing for previously unthinkable optimizations in the automation of expensive human decision making. AI is a technology deeply affecting everyone, anywhere, anytime. Still, only recently, its impact on our lives and the consequences of an AI-powered society have been brought into focus. Important questions include: What are the risks of the technology? Is AI sustainable? What technology do we want to develop? What values should the technology reflect? In this talk, I will focus on pathways to responsible- and sustainable AI by considering aspects such as fairness, explainability, privacy, environmental impact and sustainable development.

AI in the wild - a literal title-2-image translation



- " in the wild" → in real world
- data-driven AI → software agents learning from data

Outline

- Why it is important/ why now?
- Fairness-aware learning
- Explainability
- After deployment
- AI for sustainable design
- Wrapping up

Many successful applications

Recommendations

Navigation

Severe weather alerts

Automation

Great circumstances

- Data enablers
  - Web
  - Internet of things
  - Data intensive science
  - Big data
- (Intelligent) technology enablers
  - Mature DM, ML, AI
  - Deep learning
- Infrastructure enablers
  - Hardware advances
  - Software advances
- Everyone wants to join

Data deluge

AI/ML advances

Computer power

Participation

And everyone can join … The democratization of data and AI

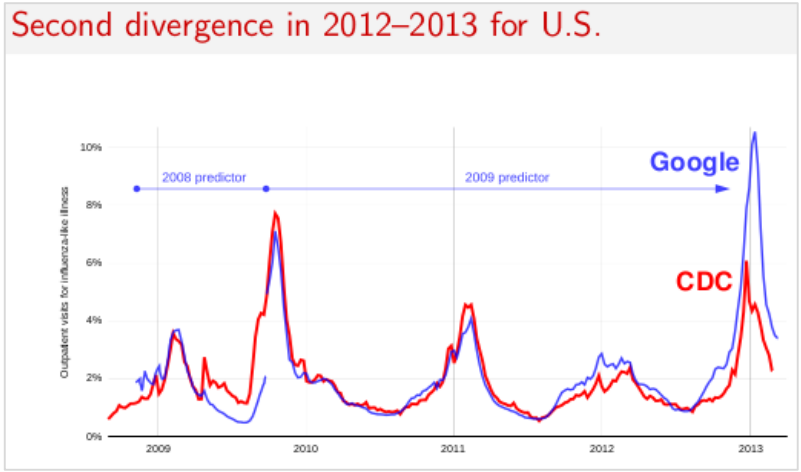
- Data democratization
  - "Data democratization is the ability for information in a digital format to be accessible to the average end user. The goal of data democratization is to allow non-specialists to be able to gather and analyze data without requiring outside help."

    Source: "*https://www.techtarget.com/whatis/definition/data-democratization*"
- AI democratization
  - "AI democratization is the spread of AI development to a wider user base that includes those without specialized knowledge of AI. The trend is being driven by large companies that are heavily invested in artificial intelligence, including IBM, Amazon, Facebook, Microsoft and Google, in the interests of furthering its development and adoption.
  - AI development has typically demanded a lot of resources including expert-level knowledge, computing power and money. AI democratization involves facilitating development by providing user-friendly resources and supports such as pre-built algorithms, intuitive interfaces and high-performance cloud computing platforms. Having those supports in place makes it feasible for in-house developers without special expertise to create their own machine learning applications and other AI software."

    Source: *https://whatis.techtarget.com/definition/AI-democratization*

Failures/bad use of the technology

- **Google's flue trends (GFT)** failure
- Main idea: using search data to predict flu
  - Intuition: when people are sick with the flu, many search for flu-related information on Google, providing almost instant signals of overall flu prevalence
  - Results: "we can accurately estimate the current level of weekly influenza activity in each region of the United States, with a reporting lag of about one day." Source
- GFT failed missing at the peak of the 2013 flu season by 140 percent.
- Why it failed?
  - Google's algorithm was quite vulnerable to overfitting to seasonal terms unrelated to the flu, like "high school basketball." Source → Correlation is not causation
  - Changes in user behavior over time were not considered
  - …



What can we learn from this failure? Are big data useless?

No, other research has demonstrated the value of big data in modeling disease spread, real time identification of emergencies etc.
But we need to implement the technology properly.

Failures/bad use of the technology

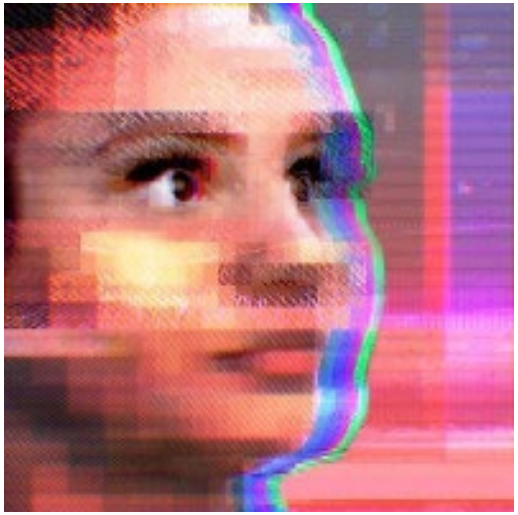
- **IBM Watson for oncology** cancelled
- Main idea: build a tool for diagnosing and treating patients (treatment recommendations)
- The program provided 'often inaccurate' and 'unsafe' treatment recommendations for patients source
- Why it failed?
  - Problem complexity
  - Training data:
    - "most of the data fed to it is hypothetical and not real patient data" source
    - "Synthetic cases allow you to treat and train Watson on a variety of patient variables and conditions that might not be present in random patient samples, but are important to treatment recommendations" source
  - …

What can we learn from this failure? Should we forget about AI in healthcare?

No, but we need to implement the technology properly.

Other important aspects: who owns the data, who is responsible for it, who can use it.

Failures/bad use of the technology



- **Microsoft's Tay bot** taken offline
  - **Tay** was an artificial intelligence chatter bot that was originally released by Microsoft Corporation via Twitter on March 23, 2016
  - **Main idea:** Tay was designed to mimic the language patterns of a 19-year-old American girl, and to learn from interacting with human users of Twitter.[7]
  - The bot began to post inflammatory, offensive, racist tweets through its Twitter account, causing Microsoft to shut down the service only 16 hours after its launch.[1]
- Why it failed?
  - According to Microsoft, this was caused by trolls who "attacked" the service as the bot made replies based on its interactions with people on Twitter.[2]

What can we learn from this failure?

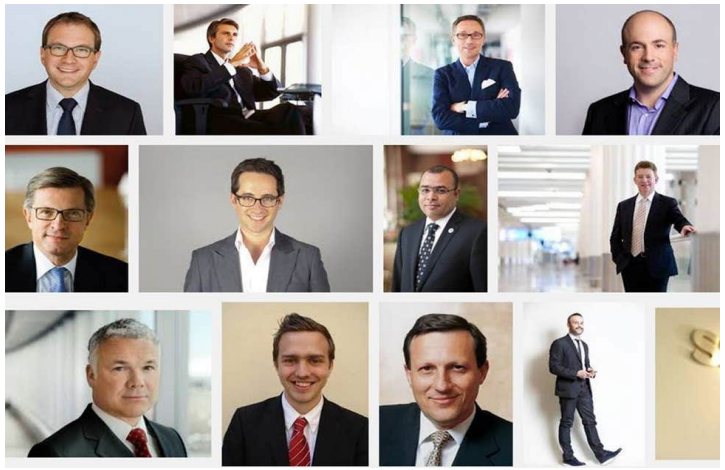
Online learning/Continual learning/Never ending learning are necessary components for strong AI.

But we should built technology that is resilient also to misuse.

And we need to embed ethical etc societal values in the technology

Biased use of the technology

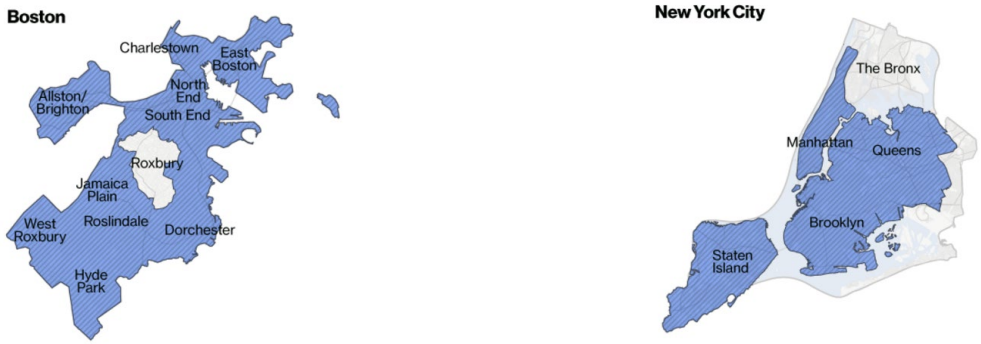
Facial recognition: State of the art visions systems (used e.g., in autonomous driving) recognize better white males than black women (*racial- and gender-bias*)



Job market: Google's AdFisher tool for serving personalized ads was found to serve significantly fewer ads for high paid jobs to women than men (*gender-bias*)

Biased use of the technology

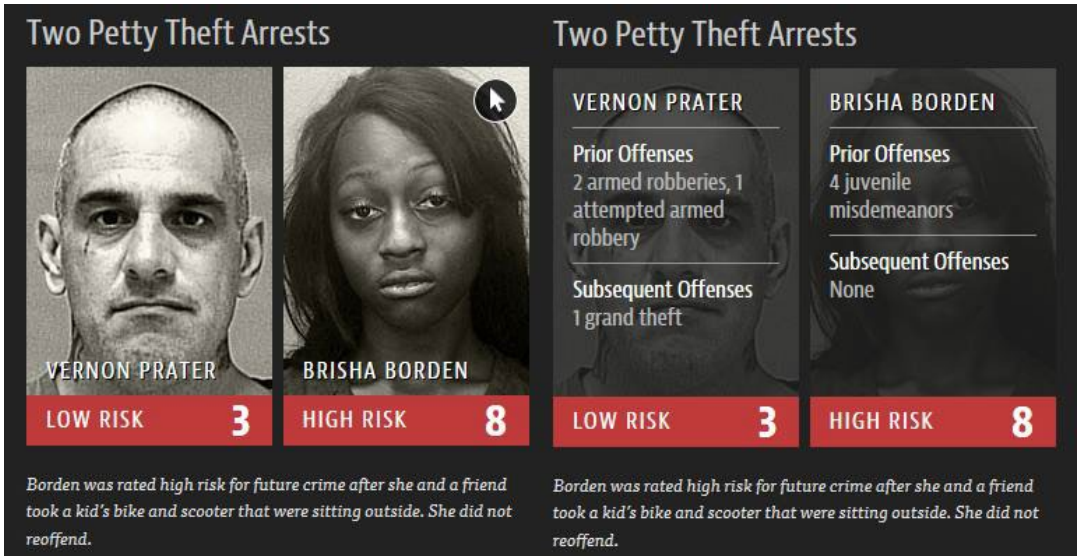
- Bloomberg analysts compared Amazon same-day delivery areas with U.S. Census Bureau data
- They found that in 6 major same-day delivery cities, the service area excludes predominantly black ZIP codes to varying degrees (*racial bias*).



Source: https://www.bloomberg.com/graphics/2016-amazon-same-day/

- Amazon claimed that race was not used in their models.

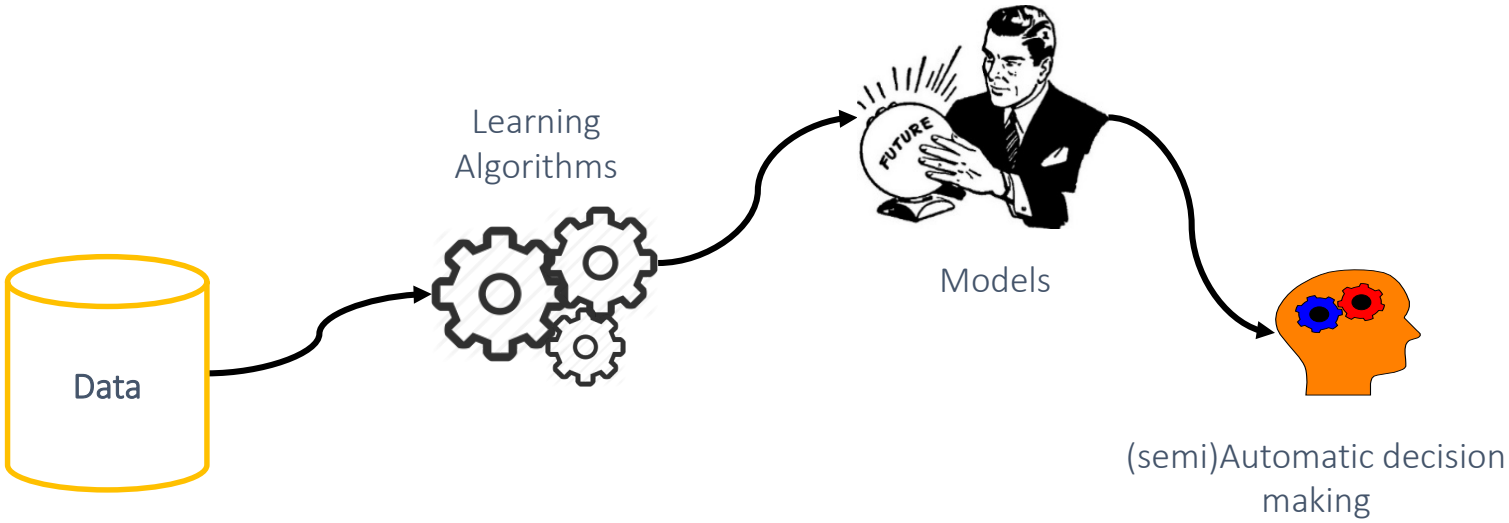
Biased use of the technology



Recidivism prediction: COMPAS tool (US) for predicting a defendant's risk of committing another crime predicted higher risks of recidivism for black defendants (and lower for white defendants) than their actual risk (*racial-bias*)

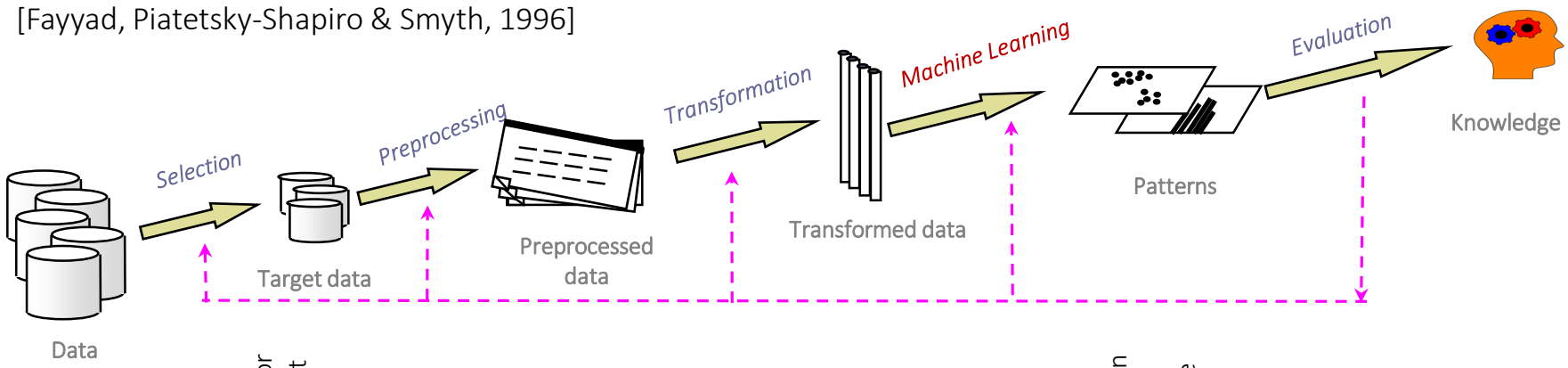
~~How machines learn~~ How we teach machines?

- ML "*gives computers the ability to learn without being explicitly programmed*" (Arthur Samuel, 1959)
- We don't codify the solution. We don't even know it!
- Data is the key & the learning algorithm



AI/ML/DS pipelines: endless choices and decisions (typically made by data scientists)

[Fayyad, Piatetsky-Shapiro & Smyth, 1996]



Selection:
- Select a relevant dataset or focus on a subset of a dataset
- File / DB/

Preprocessing/Cleaning:
- Integration of data from different data sources
- Noise removal
- Missing values

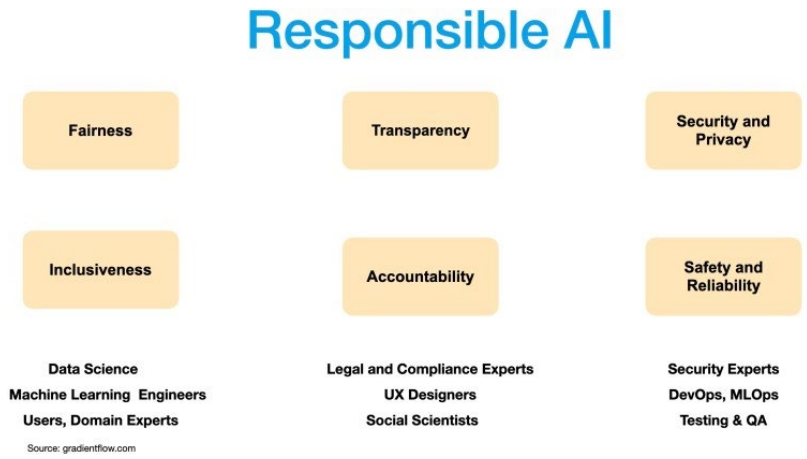
Transformation:
- Select useful features
- Feature transformation/ discretization
- Dimensionality reduction

Machine Learning:
- Search for patterns of interest

Evaluation:
- Evaluate patterns based on interestingness measures
- Statistical validation of the Models
- Visualization
- Descriptive Statistics

Moving forward

- A growing interest in principles, tools, and best practices for deploying AI ethically and responsibly.



- "Sustainable AI is a movement to foster change in the entire lifecycle of AI products (i.e. idea generation, training, re-tuning, implementation, governance) towards greater ecological integrity and social justice" (van Wynsberghe, 21)

Outline

- Why it is important/ why now?
- Fairness-aware learning
- Explainability
- After deployment
- AI for sustainable design
- Wrapping up

Moving forward: fairness-aware learning

- A young, fast evolving, multi-disciplinary research field
  - Bias/fairness/discrimination/… have been studied for long in philosophy, social sciences, law, …

- Existing fairness-aware ML approaches can be divided into three categories
  - Understanding bias
    - How bias is created in the society and enters our sociotechnical systems, is manifested in the data used by AI algorithms, and can be formalized.
  - Mitigating bias
    - Approaches that tackle bias in different stages of AI-decision making.
  - Accounting for bias
    - Approaches that account for bias proactively or retroactively.

E. Ntoutsi, P. Fafalios, U. Gadiraju, V. Iosifidis, W. Nejdl, M.-E. Vidal, S. Ruggieri, F. Turini, S. Papadopoulos, E. Krasanakis, I. Kompatsiaris, K. Kinder-Kurlanda, C. Wagner, F. Karimi, M. Fernandez, H. Alani, B. Berendt, T. Kruegel, C. Heinze, K. Broelemann, G. Kasneci, T. Tiropanis, S. Staab" *Bias in data-driven artificial intelligence systems—An introductory survey*", WIREs Data Mining and Knowledge Discovery, 2020.

NoBIAS
Artificial Intelligence without Bias

https://nobias-project.eu/
@NoBIAS_ITN

Understanding bias: Sociotechnical causes of bias

- AI-systems rely on data *generated* by humans (UGC) or *collected* via systems created by humans.

Extreme *she* occupations

1. homemaker	2. nurse	3. receptionist
4. librarian	5. socialite	6. hairdresser
7. nanny	8. bookkeeper	9. stylist
10. housekeeper	11. interior designer	12. guidance counselor

Extreme *he* occupations

1. maestro	2. skipper	3. protege
4. philosopher	5. captain	6. architect
7. financier	8. warrior	9. broadcaster
10. magician	11. figher pilot	12. boss

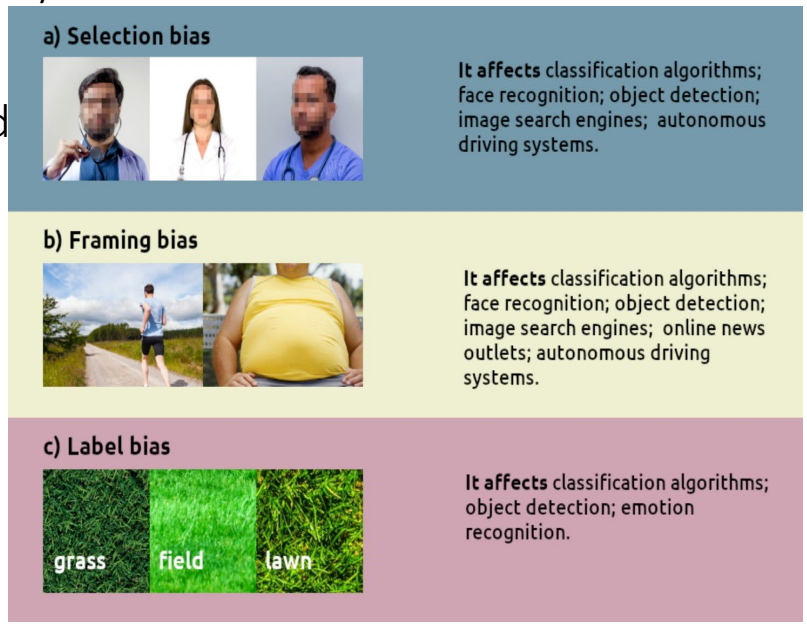
- As a result human biases
  - enter AI systems
    - e.g., gender bias in word-embeddings (Bolukbasi et al, 2016)
    - e.g., racial bias in computer vision datasets (Buolamwini and Gebru, 2018)
  - might be amplified by complex sociotechnical systems such as the Web
    - e.g., how the Web amplifies polarization[1]
  - might be amplified by feedback loops and pipelines
    - e.g., using biased word-embeddings as component of a job recommender system
  - new types of biases might be created

[1] *https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0213246*

Berendt, B., Gandon, F., Halford, S., Hall, W., Hendler, J., Kinder-Kurlanda, K., Ntoutsi, E. and Staab, S., 2021. Web Futures: Inclusive, Intelligent, Sustainable The 2020 Manifesto for Web Science. *Dagstuhl Manifestos*.

Understanding bias: How is bias manifested in data?

- Protected attributes and proxies
  - E.g., neighborhoods in U.S. cities are highly correlated with race
- Representativeness of data
  - E.g., underrepresentation of women and communities and image datasets
  - E.g., overrepresentation of black people
- Depends on data modalities
  - Text vs images vs … vs multimodal data



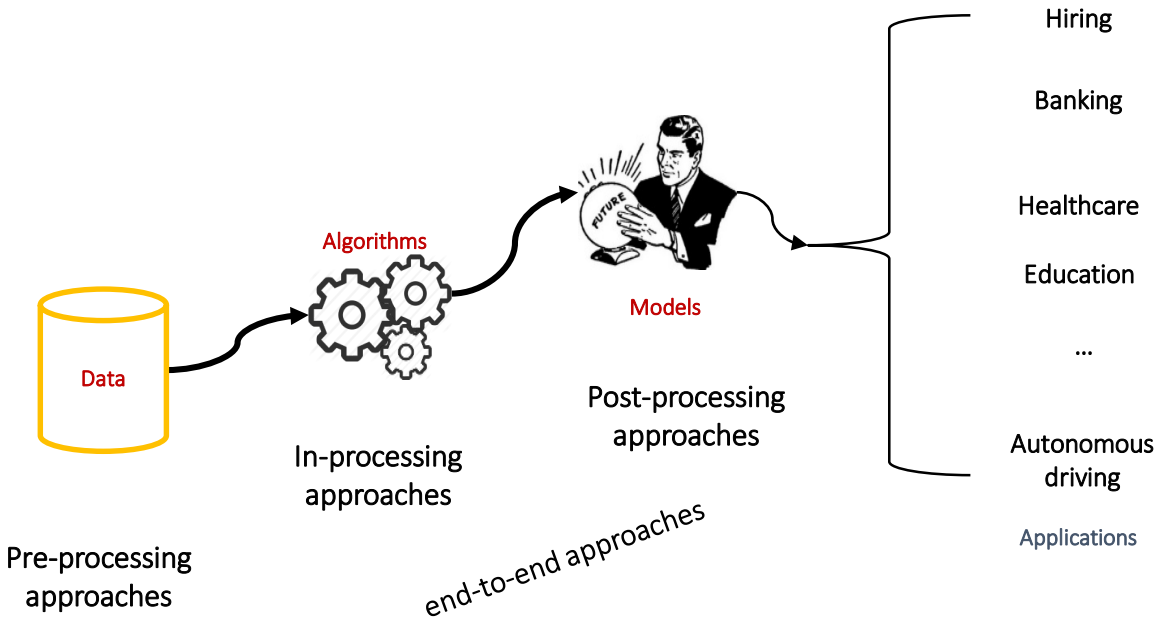
Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., & Kompatsiaris, I. (2021). A survey on bias in visual datasets. *arXiv preprint arXiv:2107.07919*.

Understanding bias: How is fairness defined?

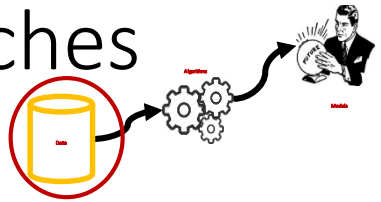
- A wide variety of fairness definitions (Verma and Rubin, 2018)
- Types of fairness measures: group fairness, individual fairness
- Group fairness: protected s (e.g., females) and non-protected $\bar{s}$ (e.g., males) groups should be treated similarly
  - Representative measures: statistical parity, equal opportunity, equalized odds, disparate mistreatment
  - Main critic: when focusing on the group less qualified members may be chosen
- Individual fairness: similar individuals should be treated similarly
  - Representative measures: counterfactual fairness
  - Main critic: it is hard to evaluate proximity of instances (M. Kim et al, NeurIPS 2018)

Mitigating bias

- Goal: tackling bias in different stages of AI-decision making

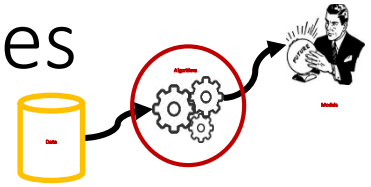


Mitigating bias: pre-processing approaches



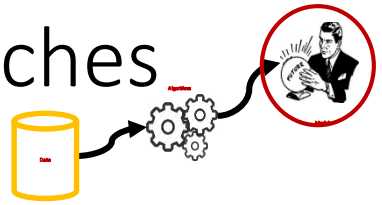
- **Intuition:** making the data “more fair” will result in a “less unfair” model
- **Idea:** balance the protected and non-protected groups in the dataset
- **Design principle:** minimal data interventions (to retain data utility for the learning task)
- Different techniques:
  - Instance selection (sampling), (Kamiran & Calders, 2010) (Kamiran & Calders, 2012)
  - Instance weighting, (Calders, Kamiran, & Pechenizkiy, 2009)
  - Instance class modification (massaging), (Kamiran & Calders, 2009),(Luong, Ruggieri, & Turini, 2011)
  - Synthetic instance generation (Iosifidis & Ntoutsi, 2018)
  - …

Mitigating bias: in-processing approaches



- **Intuition:** working directly with the algorithm allows for better control
- **Idea:** explicitly incorporate the model's discrimination behavior in the objective function
- **Design principle:** "balancing" predictive- and fairness-performance
- Different techniques:
  - Regularization (Kamiran et al, 2010),(Kamishima et al, 2012), (Dwork et al, 2012) (Zhang & Ntoutsi, 2019)
  - Constraints (Zafar et al, 2017)
  - Training on latent target labels (Krasanakis et al, 2018)
  - In-training altering of data distribution (Iosifidis & Ntoutsi, 2019)
  - …

Mitigating bias: post-processing approaches



- **Intuition:** start with predictive performance
- **Idea:** first optimize the model for predictive performance and then tune for fairness
- **Design principle:** minimal interventions (to retain model predictive performance)
- Different techniques:
  - Correct the confidence scores (Pedreschi et al, 2009), (Calders & Verwer, 2010)
  - Correct the class labels (Kamiran et al., 2010)
  - Change the decision boundary (Kamiran et al, 2018), (Hardt et al, 2016)
  - Wrap a fair classifier on top of a black-box learner (Agarwal et al, 2018)
  - …

Accounting for bias

- Algorithmic accountability refers to the assignment of responsibility for how an algorithm is created and its impact on society (Kaplan et al, 2019).
- Two categories of approaches
  - Proactive approaches:
    - Bias-aware data collection, e.g., for Web data, crowd-sourcing
    - Bias-description and modeling, e.g., via ontologies
    - ...
  - Retroactive approaches:
    - Explaining AI decisions in order to understand whether decisions are biased
    - Counterfactuals
    - …

Proactive approaches: bias-aware data collection

- For example, attempts for bias-aware visual data collection (Fabbrizzi et al, 21)
- Constructed for specific purposes, not universally bias free datasets

Paper	Year	Size	Protected attributes	Pre-existing Content	Labelling Process	Selection	Framing	Label
Buolamwini and Gebru (2018)	2018	1.2K images	Binary gender; skin tone	Yes	Human annotators; experts; additional information	•		•
Karkkainen and Joo (2021)	2021	180.5K images	Age; binary gender; race	Yes	Crowd workers	•		
Merler et al. (2019)	2019	970K images	Age; Binary gender; skin tone	Yes	Machine annotators; crowd workers	•	•	
Georgopoulos et al. (2020)	2020	41K images; 44K videos	Age; binary gender	Yes	Human annotators; machine annotators	•	•	
Barbu et al. (2019)	2019	50K images	-	No	-		•	
Wu et al. (2020) (Inclusive Benchmark)	2020	12K images	Binary gender; race	Yes	Additional information	•		
Wu et al. (2020) (Non-binary Gender Benchmark)	2020	2K images	Non-binary gender; race	Yes	Additional information	•		•
Hazirbas et al. (2021)	2021	45.1K videos	Age; Non-binary gender; skin tone	No	Human annotators; self-provided labels	•	•	•

Table 5. Summary of the datasets described in Section 4.

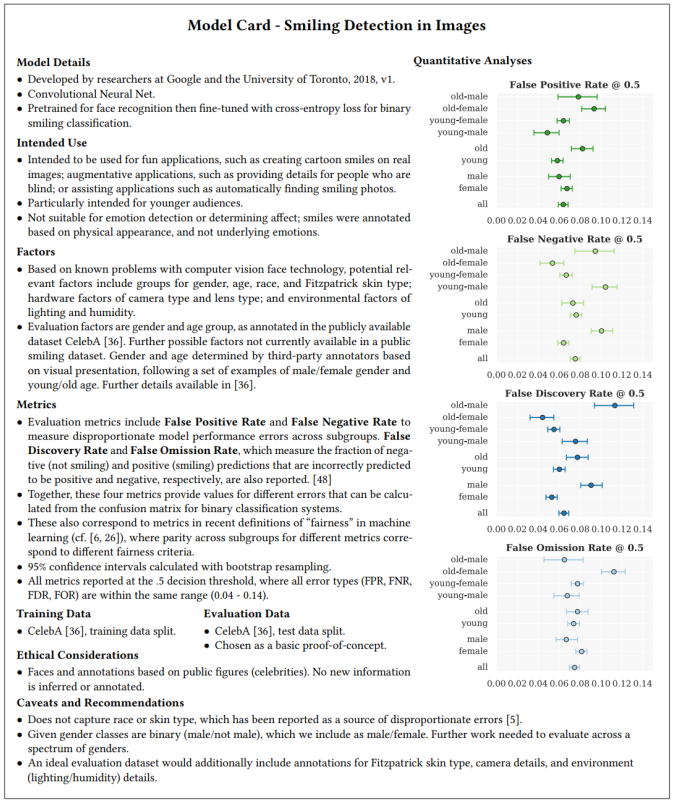
Proactive approaches: bias description and modeling

- E.g., Model cards: “aim to standardize ethical practice and reporting - allowing stakeholders to compare candidate models for deployment across not only traditional evaluation metrics but also along the axes of ethical, inclusive, and fair considerations.” (Mitchell et al, 2019)

Model Card

- **Model Details**. Basic information about the model.
  - Person or organization developing model
  - Model date
  - Model version
  - Model type
  - Information about training algorithms, parameters, fairness constraints or other applied approaches, and features
  - Paper or other resource for more information
  - Citation details
  - License
  - Where to send questions or comments about the model
- **Intended Use**. Use cases that were envisioned during development.
  - Primary intended uses
  - Primary intended users
  - Out-of-scope use cases
- **Factors**. Factors could include demographic or phenotypic groups, environmental conditions, technical attributes, or others listed in Section 4.3.
  - Relevant factors
  - Evaluation factors
- **Metrics**. Metrics should be chosen to reflect potential real-world impacts of the model.
  - Model performance measures
  - Decision thresholds
  - Variation approaches
- **Evaluation Data**. Details on the dataset(s) used for the quantitative analyses in the card.
  - Datasets
  - Motivation
  - Preprocessing
- **Training Data**. May not be possible to provide in practice. When possible, this section should mirror Evaluation Data. If such detail is not possible, minimal allowable information should be provided here, such as details of the distribution over various factors in the training datasets.
- **Quantitative Analyses**
  - Unitary results
  - Intersectional results
- **Ethical Considerations**
- **Caveats and Recommendations**

Summary of model card sections and suggested prompts for each



Model Card - Smiling Detection in Images

Model Details
- Developed by researchers at Google and the University of Toronto, 2018, v1.
- Convolutional Neural Net.
- Pretrained for face recognition then fine-tuned with cross-entropy loss for binary smiling classification.

Intended Use
- Intended to be used for fun applications, such as creating cartoon smiles on real images; augmentative applications, such as providing details for people who are blind; or assisting applications such as automatically finding smiling photos.
- Particularly intended for younger audiences.
- Not suitable for emotion detection or determining affect; smiles were annotated based on physical appearance, and not underlying emotions.

Factors
- Based on known problems with computer vision face technology, potential relevant factors include groups for gender, age, race, and Fitzpatrick skin type; hardware factors of camera type and lens type; and environmental factors of lighting and humidity.
- Evaluation factors are gender and age group, as annotated in the publicly available dataset CelebA [36]. Further possible factors not currently available in a public smiling dataset. Gender and age determined by third-party annotators based on visual presentation, following a set of examples of male/female gender and young/old age. Further details available in [36].

Metrics
- Evaluation metrics include **False Positive Rate** and **False Negative Rate** to measure disproportionate model performance errors across subgroups. **False Discovery Rate** and **False Omission Rate**, which measure the fraction of negative (not smiling) and positive (smiling) predictions that are incorrectly predicted to be positive and negative, respectively, are also reported. [48]
- Together, these four metrics provide values for different errors that can be calculated from the confusion matrix for binary classification systems.
- These also correspond to metrics in recent definitions of “fairness” in machine learning (cf. [6, 26]), where parity across subgroups for different metrics correspond to different fairness criteria.
- 95% confidence intervals calculated with bootstrap resampling.
- All metrics reported at the .5 decision threshold, where all error types (FPR, FNR, FDR, FOR) are within the same range (0.04 - 0.14).

Training Data
- CelebA [36], training data split.

Evaluation Data
- CelebA [36], test data split.
- Chosen as a basic proof-of-concept.

Ethical Considerations
- Faces and annotations based on public figures (celebrities). No new information is inferred or annotated.

Caveats and Recommendations
- Does not capture race or skin type, which has been reported as a source of disproportionate errors [5].
- Given gender classes are binary (male/not male), which we include as male/female. Further work needed to evaluate across a spectrum of genders.
- An ideal evaluation dataset would additionally include annotations for Fitzpatrick skin type, camera details, and environment (lighting/humidity) details.

Quantitative Analyses

Summing up

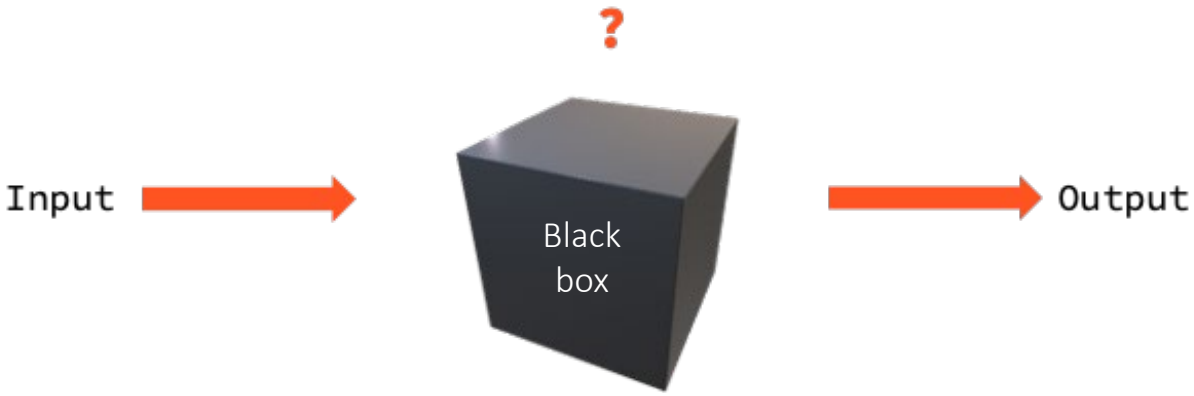
- Various “myths”/misconceptions have been debunked, e.g.
  - The myth of algorithmic objectivity: “Humans are biased, data &algorithms not”
  - Bias is caused by AI systems
- We have reached a common understanding on several aspects, e.g.
  - Technical fixes are not sufficient
  - “Fair is not fair everywhere” [Schaefer et al, 2015]
  - No single definition of fairness [Verma and Rubin, 2018]
  - Fairness is a moving target
- Still many open topics, including
  - Multi-discrimination
  - Discrimination under uncertainty
  - Discrimination for different data modalities and multimodal data
  - Evaluation aspects and long-term effects of fairness-related interventions
  - Need for benchmark datasets: “Retiring Adult: New Datasets for Fair Machine Learning”
  - Capacity building

Outline

- Why it is important/ why now?
- Fairness-aware learning
- Explainability
- After deployment
- AI for sustainable design
- Wrapping up

Moving forward: Explainability

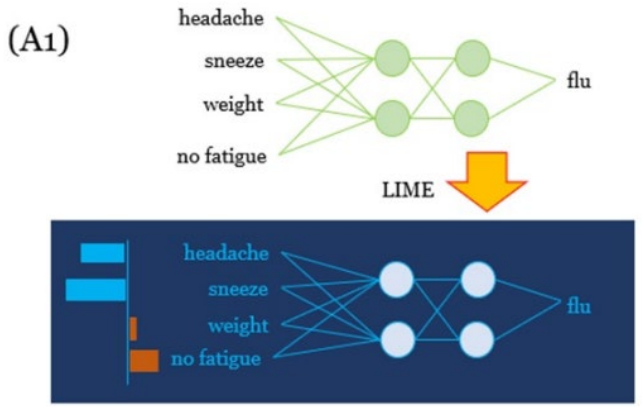
- Many AI systems nowadays are black boxes.



- This raises concerns on transparency of decision making processes (Goodman et al, 2017).
- Explainable AI (XAI): How can the internal mechanism of a (black box) model be explained in human terms?

Overview of explanation methods

- Two general categories: local vs global explanations (Guidotti et al, 2018, Das and Rad, 2020)
- Local explanations
  - Explain individual predictions
  - Representative methods:
    - Feature importance
    - Saliency maps
    - Prototype/example based
    - Counterfactuals
    - …
- Global explanations
  - Explain the complete behavior of a model
  - Representative methods
    - Model distillation
    - Representation based
    - ..



"Using LIME to generate explanation for text classification (flu). Headache and sneeze are assigned positive values. This means both factors have positive contribution to the model prediction flu. On the other hand, weight and no fatigue contribute negatively to the prediction"
Source: Tjoa and Guan, 15

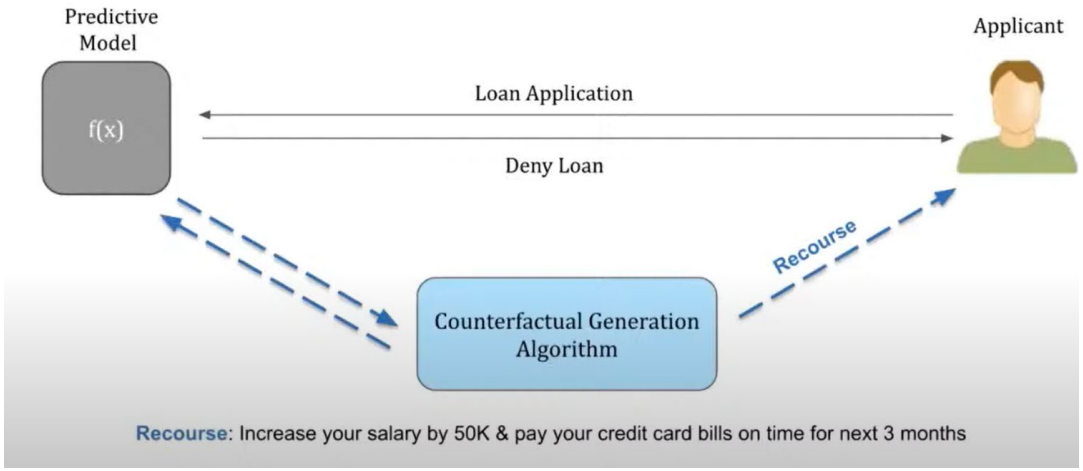
Overview of explanation methods

- Two general categories: local vs global explanations (Guidotti et al, 2018, Das and Rad, 2020)
- Local explanations
  - Explain individual predictions
  - Representative methods:
    - Feature importance
    - Saliency maps
    - Prototype/example based
    - Counterfactuals
    - …
- Global explanations
  - Explain the complete behavior of a model
  - Representative methods
    - Model distillation
    - Representation based
    - ..

Image-specific class saliency maps using gradient based attribution method
Source: Simonyan et al, 14

Explainability is a versatile tool

- For simple users, that "consume" the technology, to understand how a certain decision was made.
  - In healthcare: Why was I classified as a high risk patient for COVID?
  - In credit scoring: Why was my credit application rejected?
  - In predictive policing: Why was I selected for police inspection?
- Moreover, to get actionable insights/recommendations about the model
  - Counterfactual exlanations: What features need to be changed to flip the decision of a model? (Verma et al, 2020)



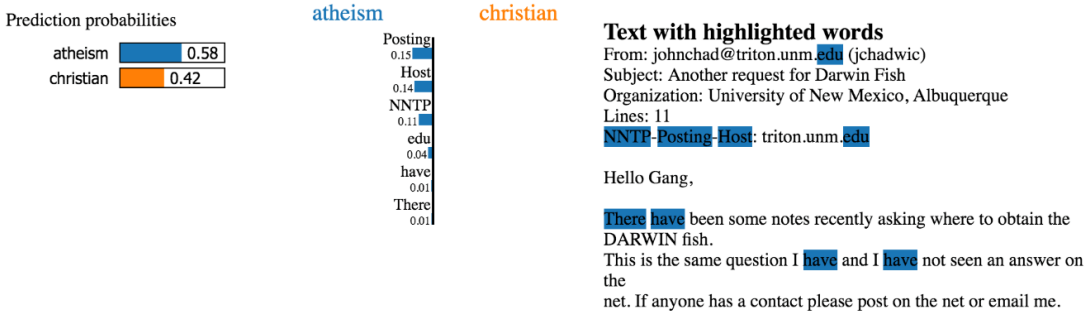
Source: (Joshi et al, 2021)

Explainability is a versatile tool

- For professionals that make decisions with (the help of) AI, moreover to ensure that decisions are correct and in accordance with legal and societal standards (e.g., no discrimination)
- For technology developers, as an inspection/debugging tool, to ensure that the technology is robust
  - Right decisions for the right reasons (Schramowski et al, 2020)
  - How to improve model performance

XAI as an inspection/debugging tool

- Explaining a text, text is predicted correctly but for the wrong reasons



- Explaining an image, predicted wrongly as Electric Guitar even though the image contains an acoustic guitar but the explanation reveals why it would confuse the two (the fretboard is very similar)

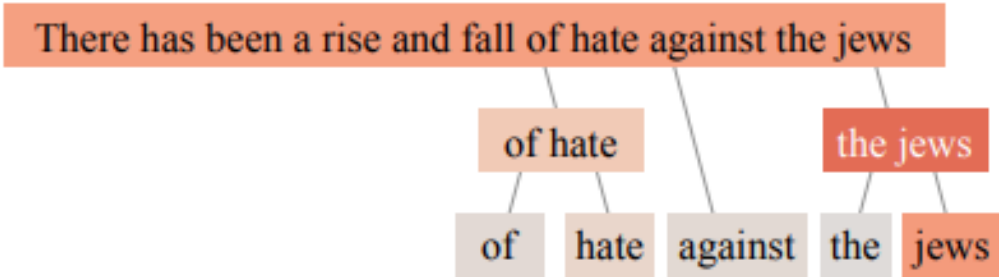
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception network, highlighting positive pixels. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Source: (Ribeiro et al, 2016)

XAI for bias detection and correction/debiasing

- Identify whether model decisions are based on group identity terms (Brendan et al, 2020)

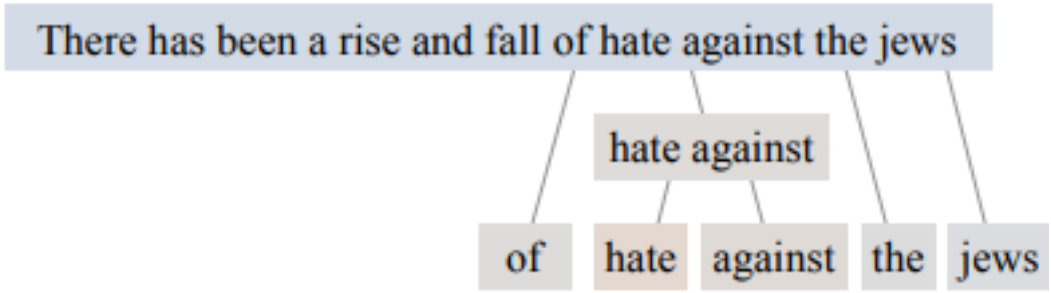


Incorrect negative attention paid to a neutral identity term
(terms with darker background are mainly responsible for the decision.)

muslim jew jews white islam blacks muslims women whites gay black democat islamic allah jewish lesbian transgender race brown woman mexican religion homosexual homosexuality africans

Examples of group identifiers

- Correct model decisions (debias), using explanation regularization



More actionability/interaction with the machine to the user

- Counterfactual explanations: What features need to be changed to flip the decision of a model? (Verma et al, 2020)
  - Assume instant materialization of changes and ignore that they may require effort and a specific order of application.
- Sequential counterfactuals: What features need to be changed and in which order to flip the decision of a model?

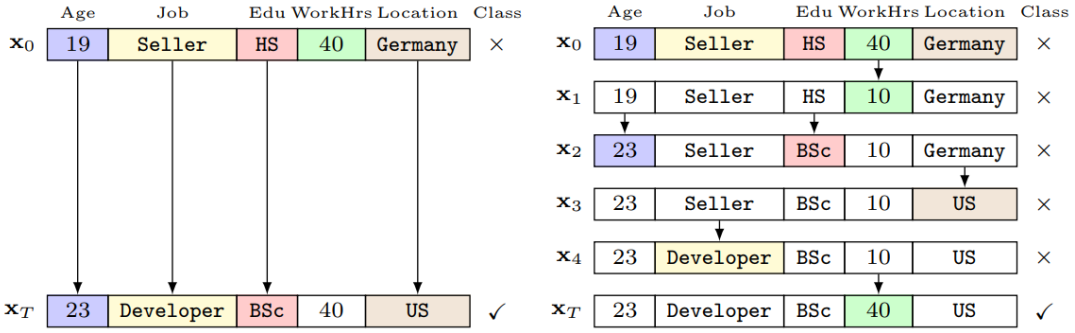


Fig. 1. The difference between traditional counterfactual generation (a) and the sequential approach (b). Although the generated counterfactual $\mathbf{x}_T$ is the same, the process and implied knowledge/information is different.

P. Naumann, E. Ntoutsi, "*Consequence-aware Sequential Counterfactual Generation*", ECMLPKDD 2021

Summing up

- XAI is an important tool to increase end-user trust in AI and allow them to gain actionable insights from the models as well as to improve model quality though inspection and correction.
- Research thus far allow us, among other things, to
  - Gain insights into the models
  - Inspect and correct models
- Still many open topics, including
  - What is a good explanation?
  - Explanations tailored to user needs
  - Considering user fatigue
  - Which explanation method to use?
  - The veracity of explanations
  - How to correct the model based on the explanations
    - In Kennedy et al, 2020 for example, the list of identity terms is given and fixed
  - Direct and indirect impact of corrections
  - Evaluation aspects
  - Building capacity

Outline

- Why it is important/ why now?
- Fairness-aware learning
- Explainability
- After deployment
- AI for sustainable design
- Wrapping up

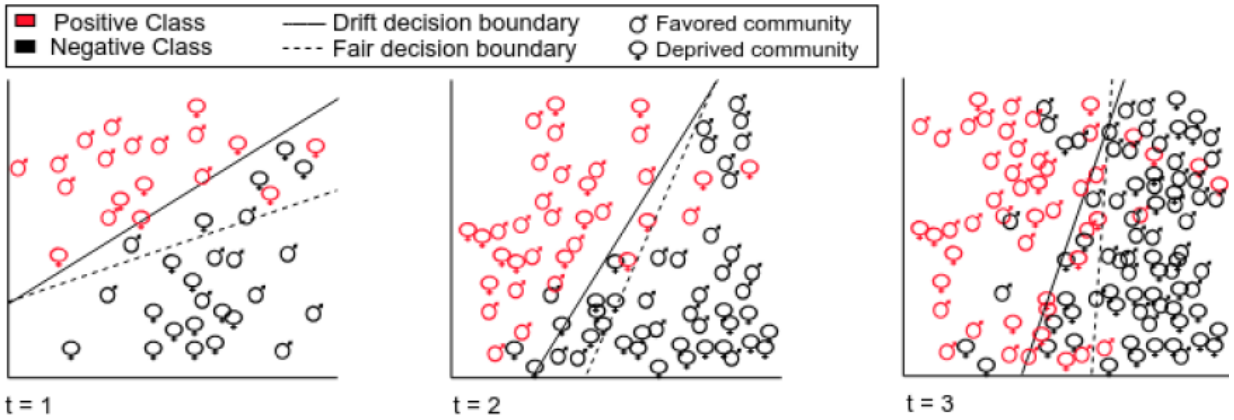
Moving forward: after deployment

- Most of the methods, stop at the model creation phase, assuming that after deployment the model will keep performing well

Happy end

Moving forward: after deployment

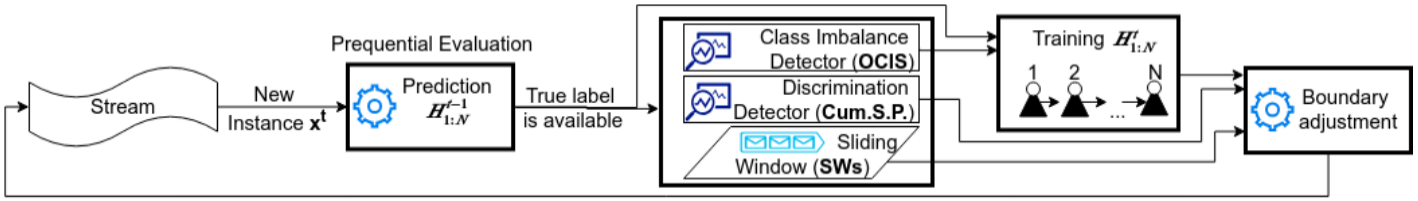
- In reality though, a well-trained model during production might not perform well after deployment due to e.g.,
  - Wrong assumptions during training (e.g., non-representative sample)
  - Assuming stationary data (but there are distribution shifts/concept drifts, so changes in the underlying data population)
- As a result, model's performance (incl. fairness) might drop



V. Iosifidis, H.T. Thi Ngoc, E. Ntoutsi, "*Fairness-enhancing interventions in stream classification*", DEXA 2019.

After deployment: fairness-aware learning

- Different lines of work: How to maintain a fair model online?
  - Monitoring of model performance over time
  - Update model as needed
    - Model update: Adding new instances but also “forgetting” outdated ones



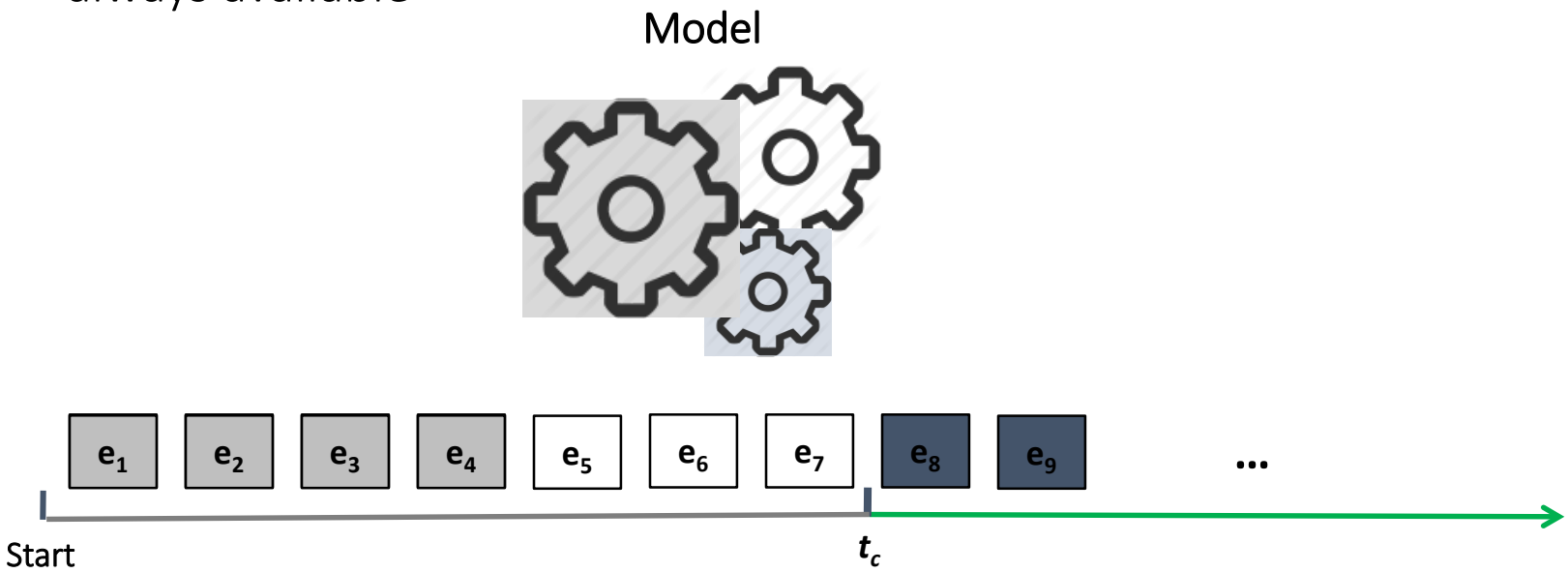
- Different approaches thus far
  - pre-processing approaches (Iosifidis et al, 19)
  - In-processing approaches (Zhang et al, 19)
  - post-processing approaches (Iosifidis et al, 20)

After deployment: fairness-aware learning

- Different lines of work: Understanding the *long-term* behaviors of deployed ML-based decision systems and their potential consequences based on simulations (D' Amour et al, 20)
- Different lines of work: Ensuring fairness up to a given horizon H (Hu et al, 22)
- …

After deployment: explainability

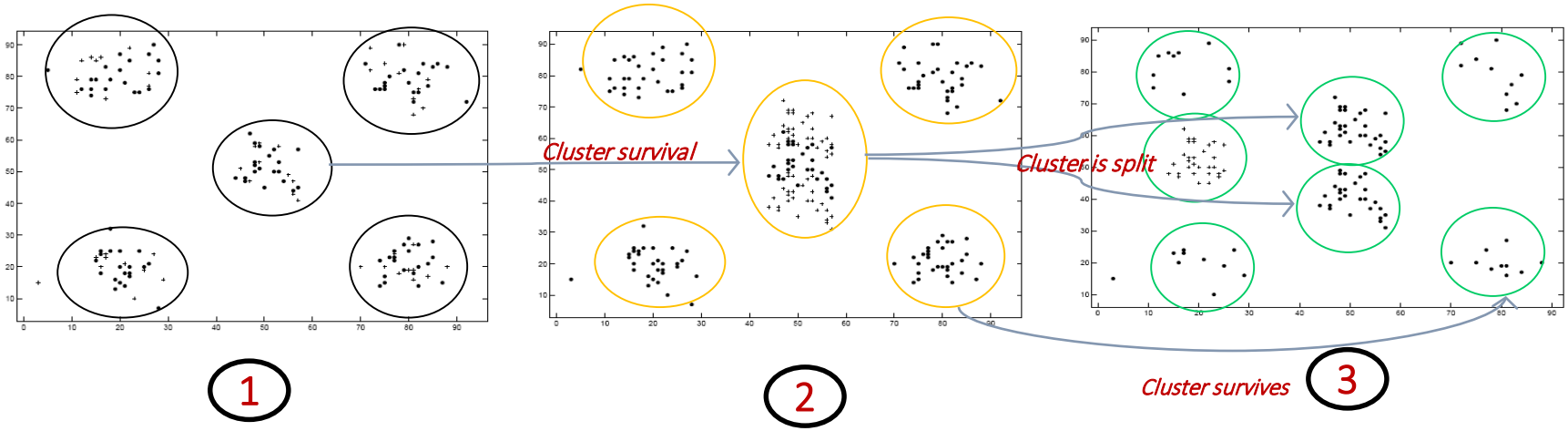
- Explaining model decisions but also explaining model changes
- Models are temporal objects and moreover the link to the data is not always available



A. Abolfazli, E. Ntoutsi, "*Drift-Aware Multi-Memory Model for Imbalanced Data Streams*", IEEE Big Data 2020

After deployment: explainability

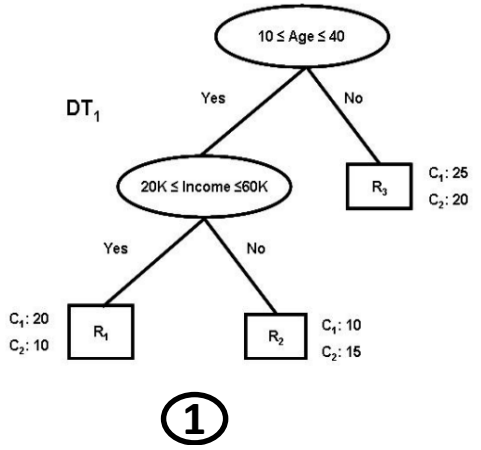
- Explaining model decisions but also explaining model changes

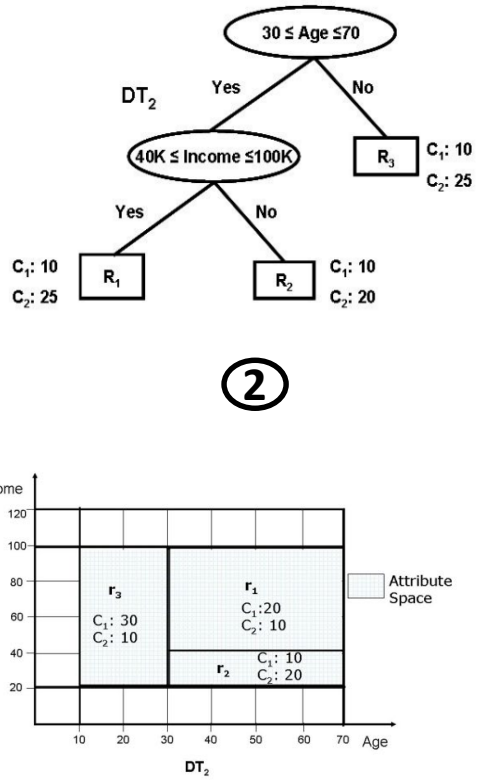


M. Spiliopoulou, E. Ntoutsi, Y. Theodoridis, R. Schult, "*MONIC: modeling and monitoring cluster transitions*", KDD 2006

After deployment: explainability

- Explaining model decisions but also explaining model changes





E. Ntoutsi, A. Kalousis, Y. Theodoridis, "*A general framework for estimating similarity of datasets and decision trees: exploring semantic similarity of decision trees*", SDM 2008

Summing up

- Most of the work stops at model creation.
- Responsibility aspects are extremely important or even more important after deployment (in the open world)
- Many exciting research directions
  - Maintaining model fairness over time
  - Explaining model decisions and model changes over time
  - Online inspection and correction of the model
  - Data provenance, a record that describes the origins and processing of data (Verder et al, 2021)
  - Constant reflection on the whole data science pipeline decisions/choices, from data collection and preprocessing to model extraction and evaluation (the devil is in the details)
  - Building capacity
  - …

Outline

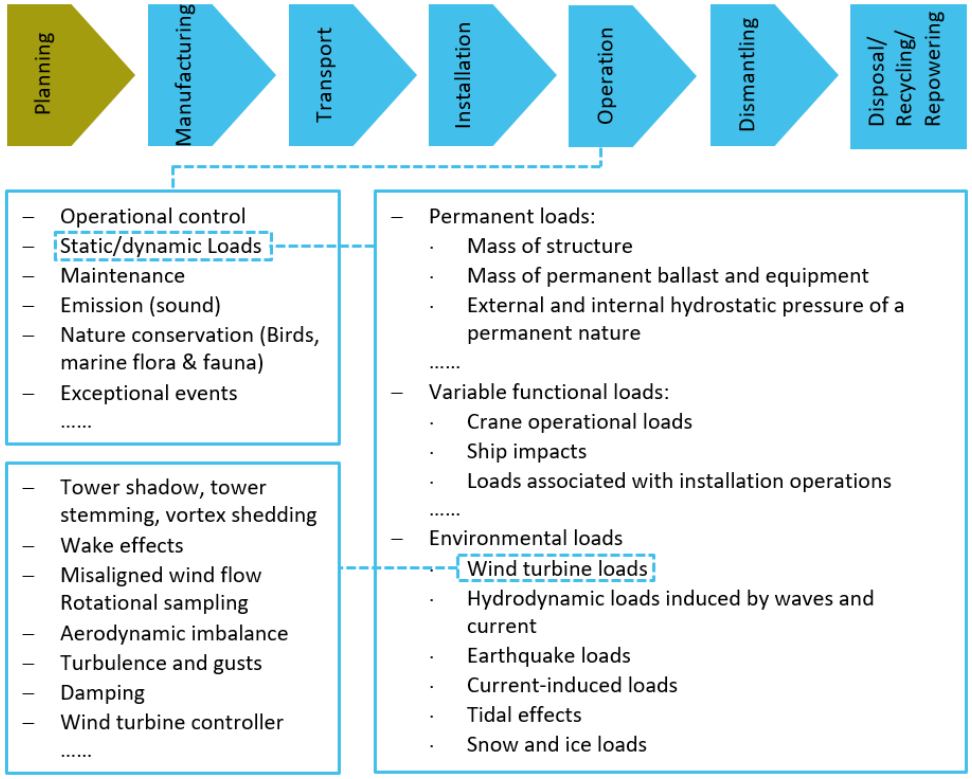
- Why it is important/ why now?
- Fairness-aware learning
- Explainability
- After deployment
- AI for sustainable design
- Wrapping up

SFB Offshore-Megastrukturen

https://www.sfb1463.uni-hannover.de/`

An example from wind turbine (WT) design

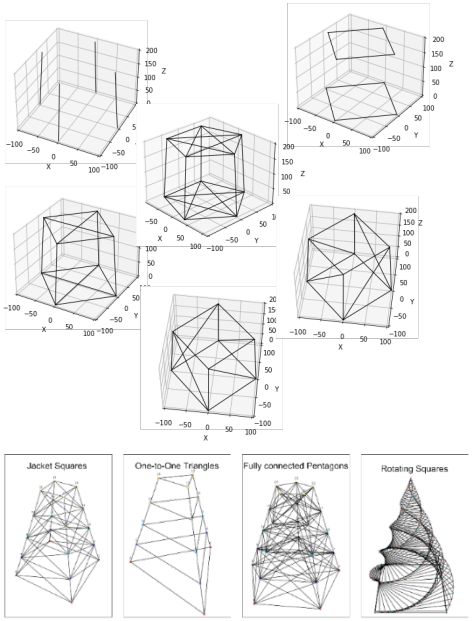
- A complex multi-step process
- Most of the existing approaches focus on optimizing the operation of a wind turbine
- Our goal is to optimize the whole life-cycle of a WT.
- A prognosis model for the prediction of the design quality with regard to the entire life cycle
- Insights on what are the parameters affecting the quality of a design →XAI
- Actionable insights/recommendations on how to improve a design → counterfactuals

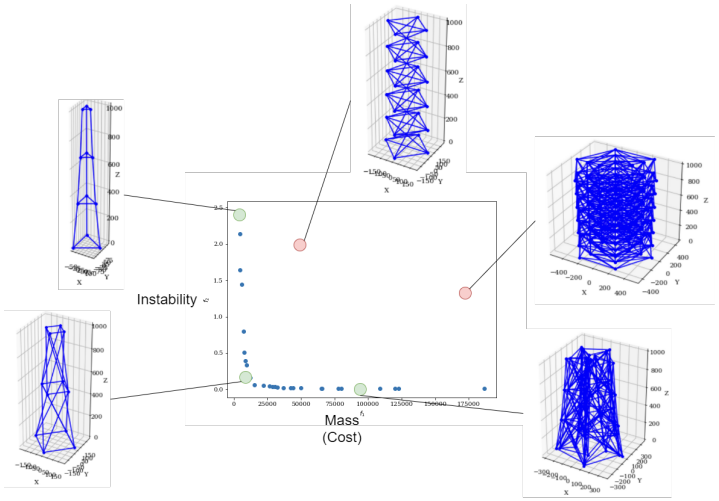


SFB Offshore-Megastrukturen

https://www.sfb1463.uni-hannover.de/

Designing wind turbines with AI





Design space

- human-defined
- learned from existing structures
- hybrid

Evaluation space

- Multiobjective (e.g., stability-related, cost-related, environmental impact-related, aesthetics, …)
- Different types of objectives: analytical models (equations), simulation models, ML-models

Other examples of generative AI

Source: https://techthelead.com/ai-turns-famous-paintings-into-photorea portraits/

Source: https://www.designboom.com/design/philippe-starck-ai-chair-kartell-interview-11-10-2020/

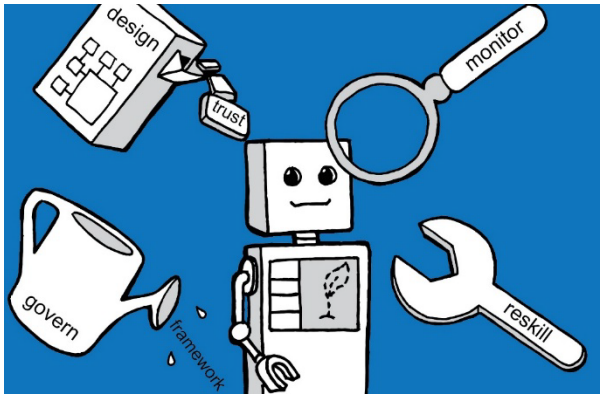
Wrapping up

https://dataresponsibly.github.io/we-are-ai/

- The technology is powerful
  - There are good applications of the technology
  - There are biased applications of the technology
  - There are bad applications of the technology
  - …
- *"New technology is not good or evil in and of itself. It's all about how people choose to use it." — David Wong (Writer)*
- However "we are AI", we can shape this technology and choose how to use it
- We means all of us ("AI is too white and male, calls for more women, minorities")
  - Beware of the data and of our design choices → Better implementation of the technology
  - Toolset: Fairness-aware learning , XAI, Interactive AI, Resilient ML, ….
- Is not only that AI needs women (diversity, in general). I believe AI is a great field for women.

Thank you for you attention!

Questions?



Feel free to contact me:
- eirini.ntoutsi@fu-berlin.de
- @entoutsi
- https://www.mi.fu-berlin.de/en/inf/groups/ag-KIML/index.html