December 2022 - Law and Ethics of artificial intelligence course

Introduction to **Responsible AI** for ML practitioners

Vivien Tran-Thien

What is Responsible AI?

An informal definition, for the purpose of this presentation

This concerns the whole lifecycle of AI systems

This covers not only the ML models but also the wider systems that embed them

The process of designing, implementing and operating AI systems in a way that prevents or minimizes the potential harms for individuals, communities, and society

This is a broad and open-ended term!

We need to consider several scales

Agenda

Three sections for a brief introduction to Responsible AI

What are the potential harms of ML use cases?

Main concerns related to the implementation of ML systems

Examples of real-life ML projects gone terribly wrong

Not included: malicious uses of ML, long term hypothetical concerns, very broad societal impacts (e.g. “future of work”)

How to identify and mitigate these potential harms along the ML project lifecycle?

Points of attention and good practices at the various stages

Roles of internal and external stakeholders

Not included: a step-by-step method that eliminates all risks if fully applied

Zoom on ML fairness

What is ML fairness? What are biases?

Some mathematical tools to detect and mitigate bias and their limitations

Not included: an in-depth and extensive review of the state-of-art techniques for ML fairness

What are the potential harms of ML use cases?

Main concerns related to the implementation of ML systems

Examples of real-life ML projects gone terribly wrong

What could go wrong?

What risks can YOU think of?

You're responsible for a new **AI project in a company**.

What could be some **unintended negative consequences** of your project?

The specific features of ML systems create new risks

These risks are amplified by the increasing adoption of ML

Features of ML systems	Potential impacts	Examples
Reliance on data Complexity and relative opacity of ML models Reduced human involvement at the design stage Automated decisions with less or no human intervention Improved performance and scalability for certain tasks	**Fairness**	

ML models may perpetuate or amplify **bias** and lead to **discrimination** against certain groups

The specific features of ML systems create new risks

These risks are amplified by the increasing adoption of ML

Features of ML systems	Potential impacts	Examples
Reliance on data	**Fairness**	
Complexity and relative opacity of ML models	**Privacy**	
Reduced human involvement at the design stage		
Automated decisions with less or no human intervention		
Improved performance and scalability for certain tasks		

Predictive models can reveal **sensitive personal information**

The specific features of ML systems create new risks

These risks are amplified by the increasing adoption of ML

Features of ML systems

Reliance on data

Complexity and relative opacity of ML models

Reduced human involvement at the design stage

Automated decisions with less or no human intervention

Improved performance and scalability for certain tasks

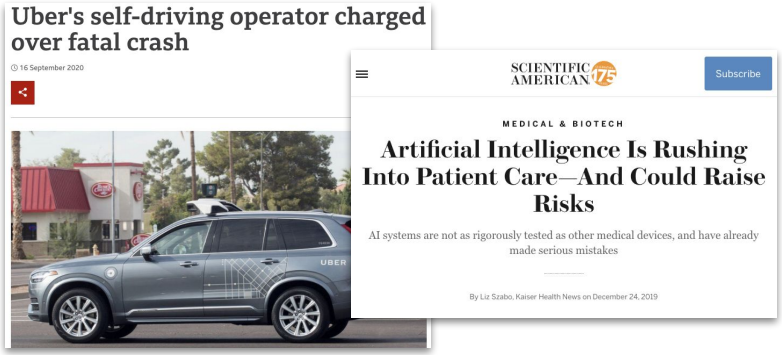
Potential impacts

Fairness

Privacy

Safety

Examples



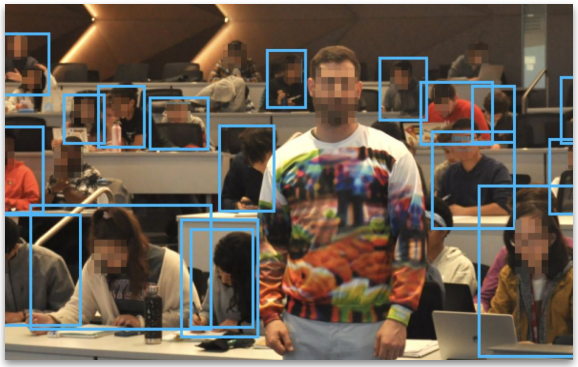
Errors have dire consequences for safety-critical applications

The specific features of ML systems create new risks

These risks are amplified by the increasing adoption of ML

Features of ML systems	Potential impacts
Reliance on data	**Fairness**
Complexity and relative opacity of ML models	**Privacy**
Reduced human involvement at the design stage	**Safety**
Automated decisions with less or no human intervention	**Security**
Improved performance and scalability for certain tasks	

Examples



Malicious actors can **fool** or **copy ML models**, and "**poison**" or reveal information about the training data

The specific features of ML systems create new risks

These risks are amplified by the increasing adoption of ML

Features of ML systems

Reliance on data

Complexity and relative opacity of ML models

Reduced human involvement at the design stage

Automated decisions with less or no human intervention

Improved performance and scalability for certain tasks

Potential impacts

Fairness

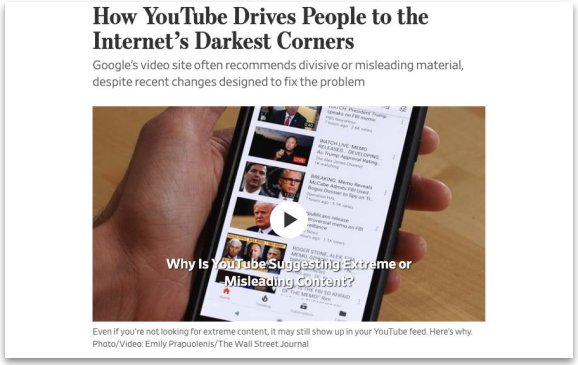
Privacy

Safety

Security

Societal impacts

Examples



Recommendation engines can promote **extremist / conspirationist content** or create **filter bubbles**

data iku

Real-life examples of ML systems gone wrong (1/5)

COMPAS: a recidivism prediction model accused of racial bias

Context

- **COMPAS** is a proprietary decision support tool created by Equivant and used by some U.S. jurisdictions.
- It assesses the **potential recidivism risk** of individuals on the basis of 137 questions (answered by these individuals or pulled from criminal records).

What went wrong

- **ProPublica**, a non-profit organization, published in 2016 an investigation showing that COMPAS is **racially biased** (even if "race" is not one of the 137 features).
- Equivant disputed these claims and argued that COMPAS' predictions are satisfying in the light of a certain fairness metric.

Takeaways

- **Not including a variable** in a model is **not enough** to prevent biased results with regard to this variable.
- Fairness is an ambivalent notion. Various **fairness metrics** may yield vastly different results.
- Can we rely on **black boxes** for **high-stakes decisions**?

Source: Machine Bias, How We Analyzed the COMPAS Recidivism Algorithm

Real-life examples of ML systems gone wrong (2/5)

Amazon renounced to use a job candidate filtering system because of gender bias

Context

- **Amazon** started developing an ML system to score job applicants (1 to 5 stars) in 2014.
- The underlying ML model was based on **50,000 terms** that had previously appeared in past candidates resumes.

What went wrong

- Reuters revealed that **the ML models developed by Amazon seemed to disadvantage women**. For example, resumes with the term “women’s” (as in “women’s chess club captain”) were penalized.
- Amazon tried to correct its ML models but ultimately renounced to use them (before Reuters’ revelations).


RETAIL

Amazon scraps a secret A.I. recruiting tool that showed bias against women

PUBLISHED WED, OCT 10 2018•6:15 AM EDT | UPDATED THU, OCT 11 2018•2:25 PM EDT

REUTERS

SHARE

Takeaways

- **Model interpretability techniques** were key to identify this fairness issue.
- **Biases in NLP systems** may be **harder to detect** because of the potentially very large number of words.
- From an external perspective, the cause of the problem is unclear: is it due to basic overfitting, biased training datasets, imbalanced training datasets, etc. ?

Source: Amazon scraps secret AI recruiting tool that showed bias against women

Real-life examples of ML systems gone wrong (3/5)

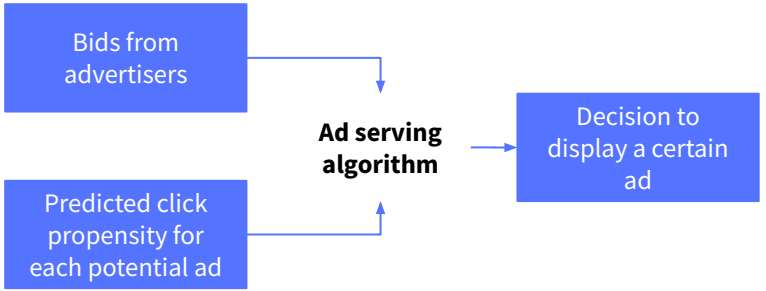
An online ad on STEM careers shown more to men than women

Context

- 2 researchers conducted a field test to determine whether an **online ad on STEM careers** would be equally shown to men and women on **Facebook**.
- They instructed the ad to be shown to both men and women and chose the same "**maximum bid per click**" for both groups.

What went wrong

- The ad ended up being **shown to men 20% more often** (and up to 45% more for the 35-44 age group).
- The ML model predicting the click propensity did not seem to be biased. The likely cause of the discrepancy seemed to be the fact that **other advertisers valued women's clicks more and outbidded the researchers**.



Takeaways

- The root cause of fairness issues in a ML system may not be the ML model itself but its **interaction with its broader environment**.
- It would have been difficult for a data scientist unfamiliar with the "economics of ad delivery" to anticipate this problem.

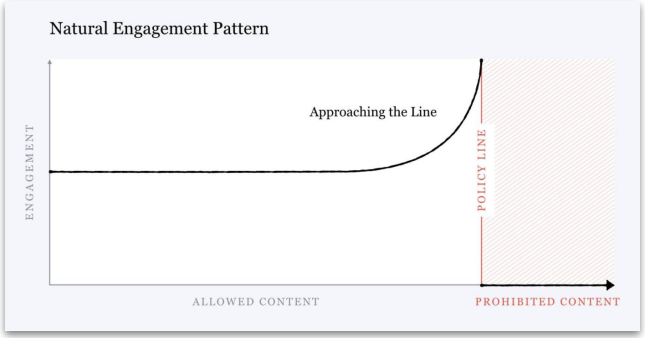
Source: Algorithmic Bias? An Empirical Study into Apparent Gender-Based Discrimination in the Display of STEM Career Ads

Real-life examples of ML systems gone wrong (4/5)

The downsides of personalized content algorithms

Context

- Because of their business models, **social media platforms** or **streaming platforms** are incentivized to maximize "**engagement**", i.e. time spent on the platform, number of videos watched...
- Several personalized content algorithms are used for this purpose: **recommendations engines**, **personalized news feed**...



What went wrong

- Maximizing engagement tend to increase the visibility of "**borderline content**" (e.g. extreme or conspirationist content).
- Ill-intentioned users can discover this trend and create purposely divisive content to generate revenues.

Takeaways

- **The ML models may not be biased** in the sense that the borderline content may indeed be more "engaging".
- The **choice of the metric to optimize** can be **very consequential**.

Source: AI Blueprint for Content Governance and Enforcement, How Facebook got addicted to spreading misinformation

Real-life examples of ML systems gone wrong (5/5)

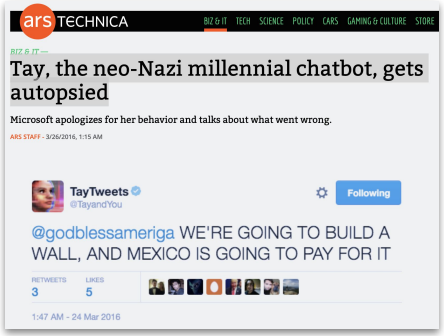
A chatbot made to write outrageous messages by Internet trolls

Context

- In 2016, Microsoft put online **Tay, a machine learning chatbot** designed to mimic a 19-year-old woman.
- It had previously successfully conducted a **similar experience in China** during which 40 million conversations happened **without major incident**.

What went wrong

- Microsoft **turned off Tay less than 24 hours later**, after Internet trolls coordinated to make Tay write offensive messages.
- The simplest trick they used was to write “repeat after me…” but some of the offensive messages also seemed to be “organic”.



Takeaways

- ML systems should be **extensively tested**, taking into account potential ill-intentioned users, especially if they can be influenced by third parties.

Source: Tay, the neo-Nazi millennial chatbot, gets autopsied

Key takeaway 1

Harms can take many shapes

- There are **several potential sources of harms**. Some may be **indirect** or occur only in **specific circumstances** and they may be subtle and **easy to overlook**.
- Responsible AI is not restricted to ML **fairness,** even if it's a key topic. Other major concerns relate to **privacy**, **safety**, **security**, or broader **societal impacts**.

Key takeaway 2

All organizations are vulnerable

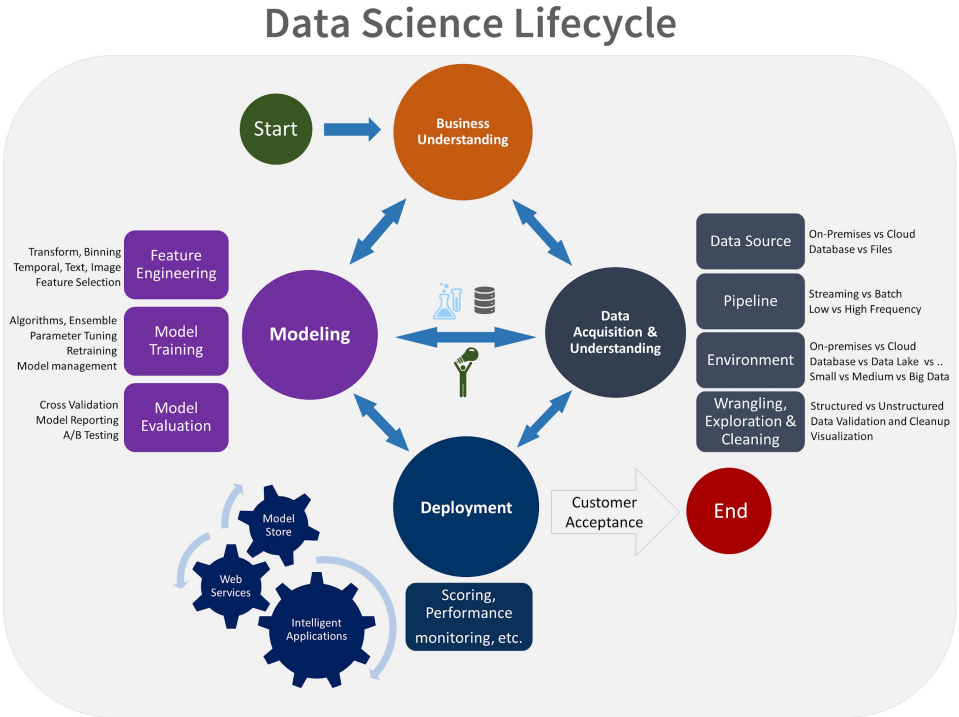
- Problems don’t only happen if you are malevolent or incompetent.
- The very public failures of some large and mature organizations are reminders to be **humble, thoughtful and rigorous**.

How to identify and mitigate these potential harms along the ML project lifecycle?

Points of attention and good practices at the various stages

Roles of internal and external stakeholders

High-level overview of the ML project lifecycle



Source: [The Team Data Science Process lifecycle](#) (Microsoft)

Step 1: Business understanding

Points of attention and good practices

Points of attention

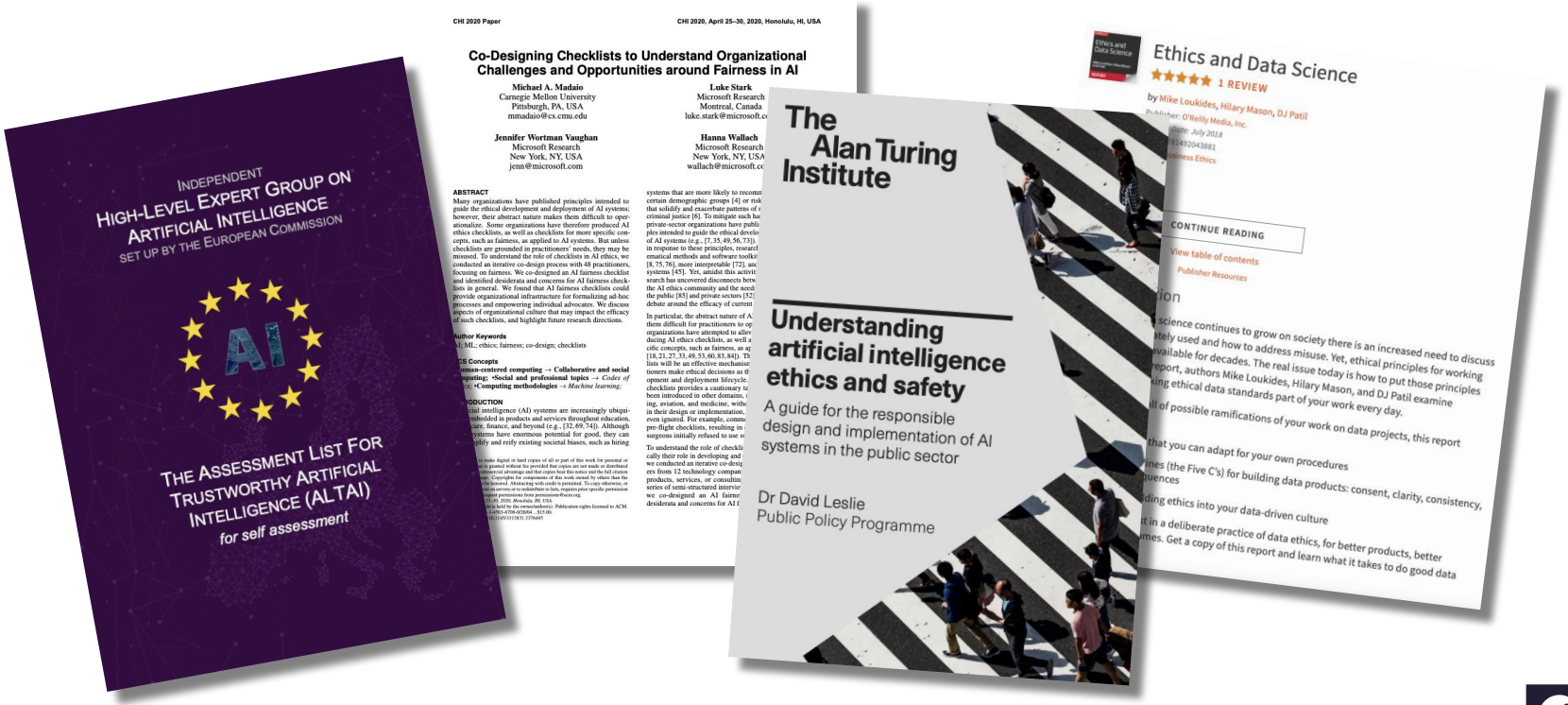
- **Success metrics** and their implications (unintended consequences, potential to be gamed)
- **Data sources** to use
- Description of the **way the ML model's predictions will be used** in practice
- High-level overview of the **impacts on the various categories of stakeholders**

Good practices

- Initiate an **assessment of the ethical aspects of your ML project** (to be refined on a regular basis) covering inter alia:
  - **Persons potentially impacted** by the project, directly or indirectly, and in normal or anomalous circumstances
  - **Nature of these impacts**
  - Possibility for adversaries to **abuse the ML system**
  - Measures foreseen to **address and monitor the ethical concerns**
- **Broadly consult** internal and external stakeholders to benefit from various expertises and perspectives

Step 1: Business understanding

Tools have been developed to facilitate the assessment of the ethical aspects of ML projects



Source: [What Are the Ethical Risks of Your AI Project?]

Exercise with 2 fictitious cases

Who is affected and what are the potential impacts?

1. Targeting restaurants for health inspections with Yelp

- **Public agency** in charge of inspecting **restaurants** to make sure they comply with food safety regulations.
- Restaurants currently chosen on the basis of customers complaints and their location
- The agency would like to use the **results of its past inspections** as well as **comments from consumers on Yelp** to **target the restaurants at a higher risk** of non compliance.

2. Generating or editing images with a text description

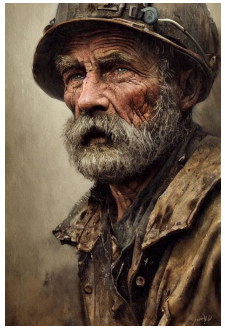
- Many proprietary and open source text-to-image models to **generate realistic or artistic images with a text description** (DALL-E, Stable Diffusion, Midjourney, Imagen…)
- **Very accessible**: simple and free tools available online
- Possibility to complete (**inpainting**, **outpainting**) or translate (**img2img**) existing images, add **new concepts or styles**

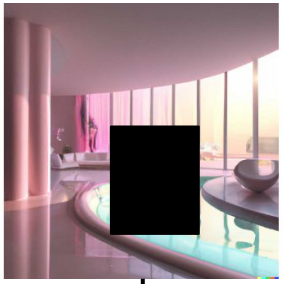
- What are the categories of stakeholders impacted?
- For each of these categories, what are potential positive or negative impacts?

→ Create teams of 5-6 students and answer the previous questions for one of the 2 cases by filling a table with the categories of stakeholders impacted as rows and two columns (“Potential positive impacts” and “Potential negative impacts”)

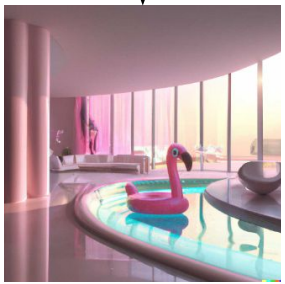
Examples of outputs of text-to-image models

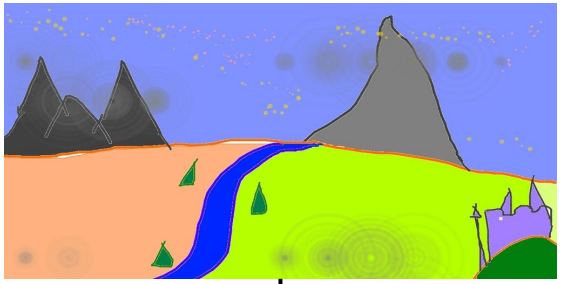
a portrait of an old coal miner in 19th century, beautiful painting with highly detailed face by greg rutkowski and magali villanueve





Inpainting





Img2img

Concept learning

Outpainting

Sources: DALL·E 2, Stable Diffusion, PromptHero, An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion

Template for the impact tables

Stakeholders	Potential positive impacts	Potential negative impacts
XXX	• … • …	• … • …
YYY	• … • …	• … • …
ZZZ	• … • …	• … • …

Case 1

Targeting restaurants for health inspections with Yelp comments

Stakeholders	Potential positive impacts	Potential negative impacts
Restaurant managers	• Reduced administrative burden for compliant restaurants	• Risks of fake reviews written by ill-intentioned competitors (to trigger inspections) • If the overall inspection results are published (e.g. through an annual report), risks of projecting a bad image to the public (because there would be more situations of non compliance detected) • Risks of being targeted because of words in the comments which are not directly connected to health problems (e.g. “nationality” of the restaurant)
Consumers	• Better protection against unhealthy food practices	• Risks of fake reviews written by restaurant managers for themselves (to avoid inspections) • Less attention for the segments of consumers (by geographical location, age group, socioeconomic status...) who don’t use social media apps like Yelp
Inspectors	• Higher professional satisfaction thanks to the improved effectiveness	• More conflictual relationships with restaurant managers who’ll be less likely to be compliant
	• Impact on the distributions of tasks (Less time on targeting? More time on administrative follow up? Impact on transportation time?)	
Taxpayers	• Better use of public resources	

Case 2

Generating or editing images with a natural language prompt

Stakeholders	Potential positive impacts	Potential negative impacts
Artists	• New creative avenues • Increased productivity	• Increased competition • Plagiarism
Art lovers	• More access to works of art • Lower barrier to create art	• Risks of uniformization or impoverishment of the art landscape
Businesses	• Easier and cheaper access to original content (e.g. advertising, marketing) • Increased productivity (e.g. image editing, design)	• Increased competition for some activities (e.g. stock photography websites)
Citizens		• Lower barrier for fake news, scams, impersonation... • Lower barrier for the creation of harmful content • Perpetuation of stereotypes • General distrust of visual evidence, reduced accountability

Step 2: Data acquisition and understanding

Points of attention and good practices

Points of attention

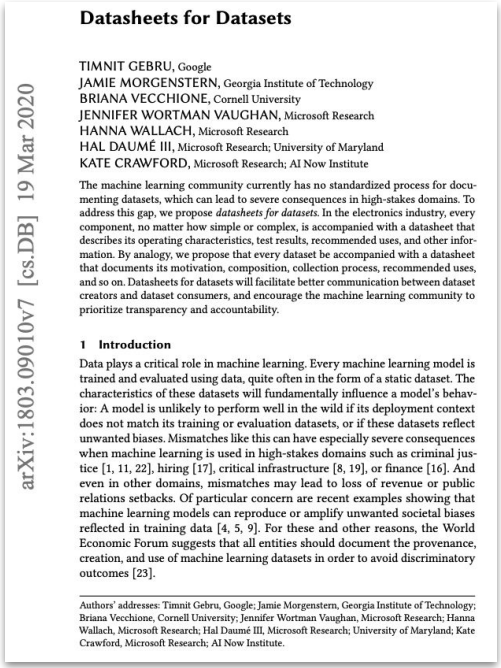
- Potential misalignment between your **data sources** and the information you really need
    - Representativeness of the data (compared to the target population)
    - Potential biases
    - Data quality issues (missing values or errors)

Good practices

- **Understand the data generation process** (data collection and labelling) with the assistance of domain experts
- **Collect and/or annotate additional data if needed**, e.g. if the current data are insufficiently representative
- **Document the datasets** in a standard format (e.g. “Datasheets for datasets”)

Step 2: Data acquisition and understanding

"Datasheets for datasets" is an article proposing a standard way to document datasets



arXiv:1803.09010v7 [cs.DB] 19 Mar 2020

Datasheets for Datasets

TIMNIT GEBRU, Google
JAMIE MORGENSTERN, Georgia Institute of Technology
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research; AI Now Institute

The machine learning community currently has no standardized process for documenting datasets, which can lead to severe consequences in high-stakes domains. To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet that describes its operating characteristics, test results, recommended uses, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. Datasheets for datasets will facilitate better communication between dataset creators and dataset consumers, and encourage the machine learning community to prioritize transparency and accountability.

1 Introduction

Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of a static dataset. The characteristics of these datasets will fundamentally influence a model's behavior: A model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted biases. Mismatches like this can have especially severe consequences when machine learning is used in high-stakes domains such as criminal justice [1, 11, 22], hiring [17], critical infrastructure [8, 19], or finance [16]. And even in other domains, mismatches may lead to loss of revenue or public relations setbacks. Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in training data [4, 5, 9]. For these and other reasons, the World Economic Forum suggests that all entities should document the provenance, creation, and use of machine learning datasets in order to avoid discriminatory outcomes [23].

Authors' addresses: Timnit Gebru, Google; Jamie Morgenstern, Georgia Institute of Technology; Briana Vecchione, Cornell University; Jennifer Wortman Vaughan, Microsoft Research; Hanna Wallach, Microsoft Research; Hal Daumé III, Microsoft Research; University of Maryland; Kate Crawford, Microsoft Research; AI Now Institute.

Motivation

For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

Labeled Faces in the Wild was created to provide images that can be used to study face recognition in the unconstrained setting where image characteristics (such as pose, illumination, resolution, focus), subject demographic makeup (such as age, gender, race) or appearance (such as hairstyle, makeup, clothing) cannot be controlled. The dataset was created for the specific task of pair matching: given a pair of images each containing a face, determine whether or not the images are of the same person.[1]

Who created this dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?

The initial version of the dataset was created by Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, most of whom were researchers at the University of Massachusetts Amherst at the time of the dataset's release in 2007.

Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.

The construction of the LFW database was supported by a United States National Science Foundation CAREER Award.

Any other comments?

Composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)? Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.

Each instance is a pair of images labeled with the name of the person in the image. Some images contain more than one face. The labeled face is the one containing the central pixel of the image—other faces should be ignored as "background".

How many instances are there in total (of each type, if appropriate)?

The dataset consists of 13,233 face images in total of 5749 unique individuals. 1680 of these subjects have two or more images and 4069 have single ones.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

[1]All information in this datasheet is taken from one of five sources. Any errors that were introduced from these sources are our fault.
Original paper: http://www.cs.cornell.edu/people/pabo/movie-review-data/. LFW survey: http://vis-www.cs.umass.edu/lfw/lfw.pdf. Paper measuring LFW demographic characteristics: http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation.MSUTechReport2014.pdf. LFW website: http://vis-www.cs.umass.edu/lfw/.

The dataset does not contain all possible instances. There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

What data does each instance consist of? "Raw" data (e.g., unprocessed text or images)or features? In either case, please provide a description.

Each instance contains a pair of images that are 250 by 250 pixels in JPEG 2.0 format.

Is there a label or target associated with each instance? If so, please provide a description.

Each image is accompanied by a label indicating the name of the person in the image.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

Everything is included in the dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

There are no known relationships between instances except for the fact that they are all individuals who appeared in news sources on line, and some individuals appear in multiple pairs.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

The dataset comes with specified train/test splits such that none of the people in the training split are in the test split and vice versa. The data is split into two views, View 1 and View 2. View 1 consists of a training subset (pairsDevTrain.txt) with 1100 pairs of matched and 1100 pairs of mismatched images, and a test subset (pairsDevTest.txt) with 500 pairs of matched and mismatched images. Practitioners can train an algorithm on the training set and test on the test set, repeating as often as necessary. Final performance results should be reported on View 2 which consists of 10 subsets of the dataset. View 2 should only be used to test the performance of the final model. We recommend reporting performance on View 2 by using leave-one-out cross validation, performing 10 experiments. That is, in each experiment, 9 subsets should be used as a training set and the 10^{th} subset should be used for testing. At a minimum, we recommend reporting the **estimated mean accuracy,** $\hat{\mu}$ and the **standard error of the mean:** S_E for View 2.
$\hat{\mu}$ is given by:

$$\hat{\mu} = \frac{\sum_{i=1}^{10} p_i}{10} \quad (1)$$

where p_i is the percentage of correct classifications on View 2 using subset i for testing. S_E is given as:

$$S_E = \frac{\hat{\sigma}}{\sqrt{10}} \quad (2)$$

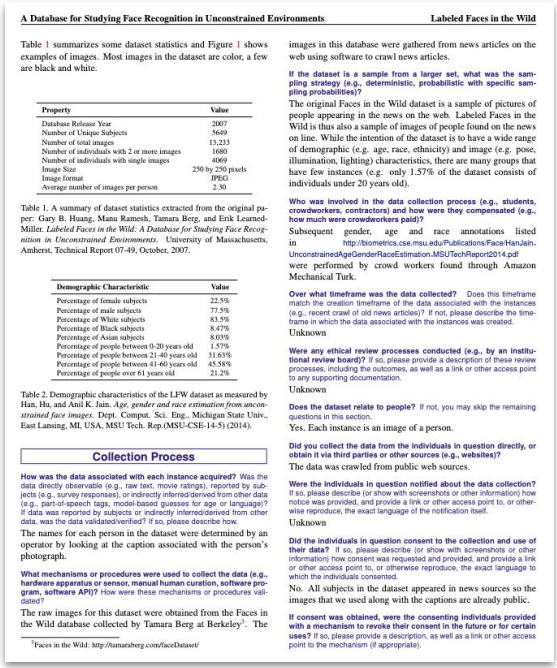


Table 1 summarizes some dataset statistics and Figure 1 shows examples of images. Most images in the dataset are color, a few are black and white.

Property	Value
Database Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%
Percentage of people between 0-20 years old	1.57%
Percentage of people between 21-40 years old	31.63%
Percentage of people between 41-60 years old	45.58%
Percentage of people over 61 years old	21.2%

Table 2. Demographic characteristics of the LFW dataset as measured by Han, Hu, and Anil K. Jain. *Age, gender and race estimation from unconstrained face images.* Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5) (2014).

Collection Process

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

The names for each person in the dataset were determined by an operator by looking at the caption associated with the person's photograph.

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

The raw images for this dataset were obtained from the Faces in the Wild database collected by Tamara Berg at Berkeley[3]. The images in this database were gathered from news articles on the web using software to crawl news articles.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

The original Faces in the Wild dataset is a sample of pictures of people appearing in the news on the web. Labeled Faces in the Wild is thus also a sample of images of people found on the news on line. While the intention of the dataset is to have a wide range of demographic (e.g. age, race, ethnicity) and image (e.g. pose, illumination, lighting) characteristics, there are many groups that have few instances (e.g. only 1.57% of the dataset consists of individuals under 20 years old).

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

Subsequent gender, age and race annotations listed in http://biometrics.cse.msu.edu/Publications/Face/HanJain_UnconstrainedAgeGenderRaceEstimation.MSUTechReport2014.pdf were performed by crowd workers found through Amazon Mechanical Turk.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

Unknown

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.

Unknown

Does the dataset relate to people? If not, you may skip the remaining questions in this section.

Yes. Each instance is an image of a person.

Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?

The data was crawled from public web sources.

Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.

Unknown

Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

No. All subjects in the dataset appeared in news sources so the images that we used along with the captions are already public.

If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).

[3]Faces in the Wild: http://tamaraberg.com/faceDataset/

Source: Datasheets for Datasets

Step 3: Modeling

Points of attention and good practices

Points of attention

- The **representativeness of the test set**
- The **errors** being made by the model
- The **situations in which the predictive performance is significantly degraded**
- The fact that the **contribution of the input features** corresponds to the domain experts' intuitions

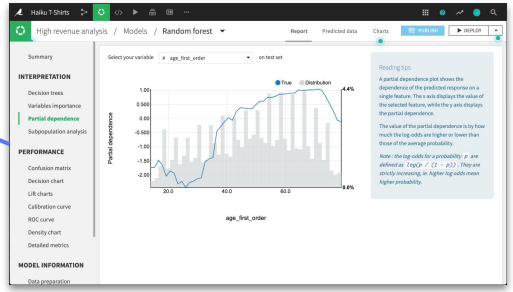
Good practices

- **Use ML fairness techniques** to detect and mitigate bias
- **Analyze errors**, the **predictive performance for various slices of the dataset**, the **contribution of the input features** with the help of a domain expert and interpretability techniques
- **Document the models** in a standard format (e.g. "Model cards for model reporting")

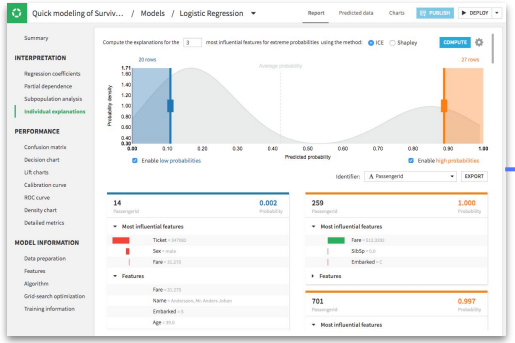
Step 3: Modeling

Various techniques and tools to analyze and documents ML models

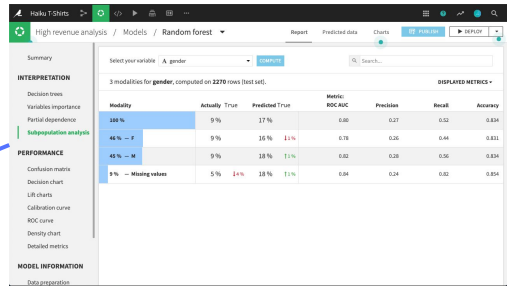
Model-level interpretability



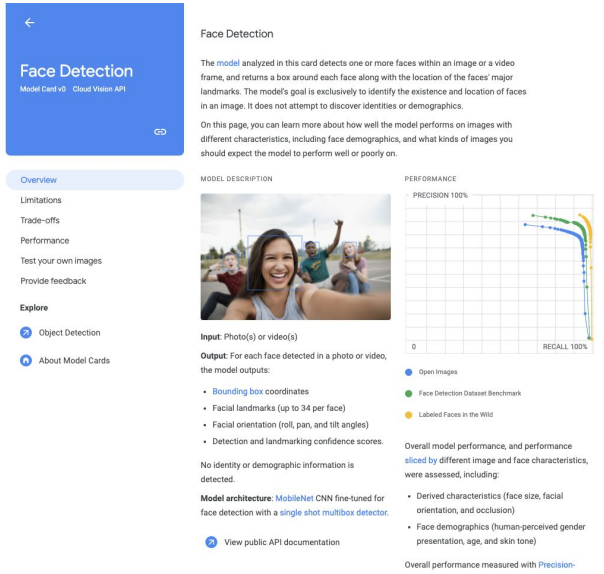
Prediction-level interpretability



Subpopulation analysis



Model cards, a standard way to document models



Source: [Model Cards for Model Reporting]

data iku

Step 4: Deployment

Points of attention and good practices

Points of attention

- If applicable (for example in the context of a decision support tool), the **behavior of human operators** that use the ML predictions. In particular: **automation bias**
- The **interactions of the ML model with its broader environment** (for example, the IT system in which it is embedded). In particular: feedback loops, consequences of anomalies or erroneous predictions...
- The **evolution** of the ML system and its environment **over time**

Good practices

- If applicable, train the **users**, adapt their **processes** and provide them with **tools** to facilitate their decisions
- **Monitor the potential negative impacts** once the model is deployed
- **Be ready, in case of errors or complaints**, to audit your ML system, provide explanations and take remedial action

Organization and governance

Points of attention and good practices

Points of attention

- **Responsible AI culture**, giving proper attention to ethical aspects of ML projects and encouraging a questioning attitude
- **Open** and **multidisciplinary approach** to benefit from a variety of perspectives
- **Accountability** for decision makers and project team members

Good practices

- Raise **awareness** about Responsible AI among internal stakeholders, in particular decision makers, and **upskill** project team members
- Define **roles** and **responsibilities**
- Adapt the **processes and methods** at each stage of the ML project lifecycle to take into consideration Responsible AI considerations
- Make the organization's **commitment to Responsible AI** explicit and give the possibility to everyone involved in ML projects to safely express doubts and concerns

Internal and external stakeholders to involve

Responsible AI concerns should be considered from a variety of perspectives

Project team members and decision makers

- **Data scientists**: analyzing datasets, training, checking and adjusting ML models, definition of the metrics to monitor in-production ML models
- **Domain experts**: providing business insights (processes, data sources, risks...)
- **Legal experts**: checking compliance with regulations and laws
- **Executives**: deciding whether the risk mitigation measures are sufficient, authorizing the operationalization of AI systems

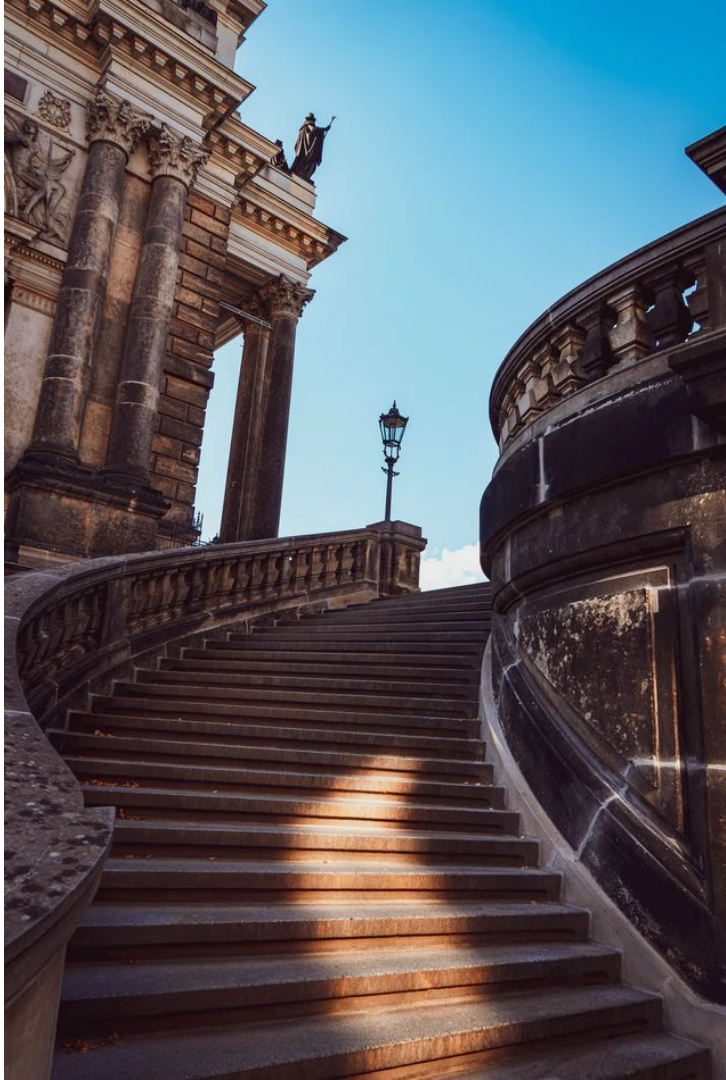
Other stakeholders

- **Users**
- Representatives of the **persons impacted by the ML system**
- **Experts**
- **NGOs**
- **Regulatory authorities**
- ...

Key takeaway 3

Responsible AI concerns the whole ML lifecycle

- Responsible AI concerns should be taken into consideration at **all stages of the ML project lifecycle**.
- Even when an ML system has been deployed, certain Responsible AI concerns need to be **monitored** and **periodically reassessed**.



Key takeaway 4

It’s a team sport

- **Technical, legal and domain expertises** are required.
- Various perspectives, from both **internal and external stakeholders**, should also be incorporated.
- **Executives should be involved for major decisions**. It isn’t up to the data scientists to decide what is acceptable or not for their organizations.
- Organizing the dialogue between the various stakeholders and facilitating the decisions of executives are essential and challenging tasks.

Zoom on ML fairness

What is ML fairness? What are biases?

Some mathematical tools to detect and mitigate bias and their limitations

data iku

What is fairness?

Well, it depends...

On which aspect? **What groups?**

Unfairness as *“Any case where AI/ML systems perform differently for different groups in ways that may be considered undesirable”*

Considered undesirable by whom?

Source: Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?

What is bias (in the context of Responsible AI)?

Bias is an ambiguous and overloaded term in AI

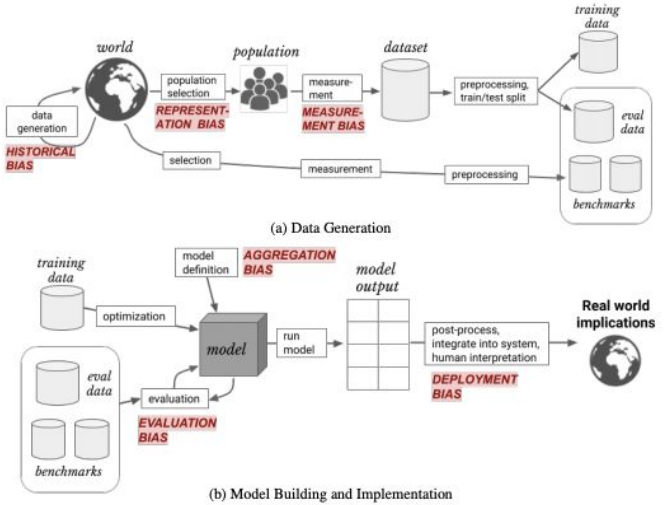
Two definitions in Google's ML Glossary

Bias (ethics/fairness)

- **Stereotyping**, **prejudice** or **favoritism** towards some things, people, or groups over others
- **Systematic error** introduced by a **sampling** or **reporting** procedure

(not to be confused with "bias" as in "bias-variance decomposition", "inductive bias" or in "bias term")

Typology of biases



Sources: Google's Machine Learning Glossary, A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

Group fairness metrics (1/3)

The most common approach to measure bias

(Simplified) formulation in the binary case

- **X**, categorical and/or continuous **features**
  E.g. diploma, years of experience...
- **A**, categorical "**sensitive**" (or "protected") **attribute**
  E.g. gender
- **Y**, binary **target** variable
  E.g. qualified for the job offer
- $\hat{Y}(X)$, **prediction** provided by the model to assess
  E.g. selected by the filtering algorithm

The model is considered fair if the **probabilities** of some events determined by the values of the **target** and the **prediction** are the same for all possible values of the **protected attribute**.

This requires only the **confusion matrices** for the various values of A. Knowing X is not required

Examples of metrics

- **Demographic parity**: P(predicted positive)
- **Equal opportunity**: P(predicted positive | positive)
- **Equalized odds**: P(predicted positive | positive), P(predicted positive | negative)
- **Predictive parity**: P(positive | predicted positive), P(positive | predicted negative)
- **Positive predictive parity**: P(positive | predicted positive)

Group fairness metrics (2/3)

The most common approach to measure bias

Group A		Prediction	
		0	**1**
Reality	**0**	30	30
	1	20	20

Group B		Prediction	
		0	**1**
Reality	**0**	25	25
	1	25	25

P(predicted positive) = **50%** — ✔ **Demographic parity** — P(predicted positive) = **50%**

P(predicted positive | positive) = **50%** — ✔ **Equalized odds** — P(predicted positive | positive) = **50%**

P(predicted positive | negative) = **50%** — ✔ **Equal opportunity** — P(predicted positive | negative) = **50%**

P(positive | predicted positive) = **40%** — X **Predictive parity** — P(positive | predicted positive) = **50%**

P(positive | predicted negative) = **40%** — X **Positive predictive parity** — P(positive | predicted negative) = **50%**

Group fairness metrics (3/3)

The most common approach to measure bias

Black defendants		COMPAS score	
		Low	High
Recidivated?	No	990	805
	Yes	532	1369

White defendants		COMPAS score	
		Low	High
Recidivated?	No	1139	349
	Yes	461	505

Accuracy = **64%**

P(high | not recidivated) = **45%**

P(low | recidivated) = **28%**

P(recidivated | high) = **63%**

P(not recidivated | low) = **65%**

X Equal opportunity

~ ✔ Predictive parity

~ ✔ Positive predictive parity

Accuracy = **67%**

P(high | not recidivated) = **23%**

P(low | recidivated) = **48%**

P(recidivated | high) = **59%**

P(not recidivated | low) = **71%**

Source: How We Analyzed the COMPAS Recidivism Algorithm

What fairness metric makes intuitive sense to you in this scenario?

Let's vote!

Scenario

- You are developing a ML model for a **job applicant filtering** system and you want to make sure your model is fair for the candidates.
- There are **two groups of candidates**. Being part of one or the other group is irrelevant in itself for the job performance (but it can be correlated with factors relevant for job performance). The candidates are either "qualified" or "not qualified".
- The candidates **selected** by the ML model will be **further assessed** before a hiring manager takes a final decision.

Intuitively, which fairness metric would make sense to you in this scenario?

1. **Demographic parity**: P(selected)
2. **Equal opportunity**: P(selected | qualified)
3. **Equalized odds**: P(selected | qualified), P(selected | not qualified)
4. **Predictive parity**: P(qualified | selected), P(qualified | not selected)
5. **Positive predictive parity**: P(qualified | selected)
6. **None!** ML models shouldn't be used for HR decisions

Group fairness isn't the only mathematical formulation of fairness

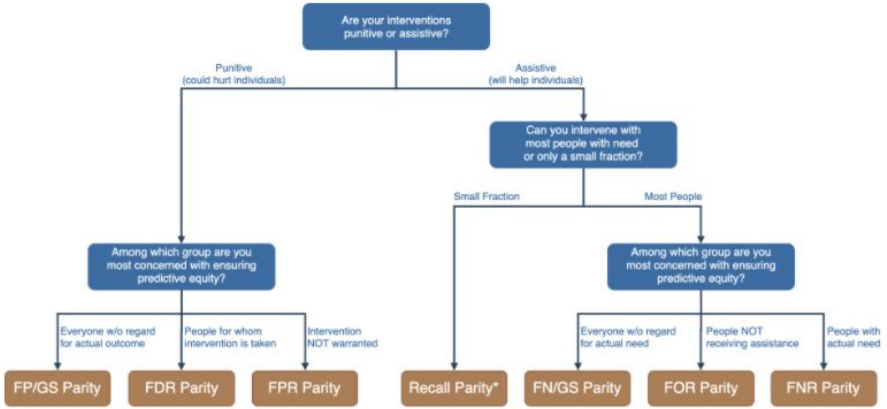
Advantages, drawbacks and alternatives

Advantages

- **Simple to understand and compute** with just the sensitive attribute, the ground truth and the prediction

Drawbacks

- **Many different metrics** that lead to significantly different outcomes and that need be carefully chosen for each use case
- Metrics sometimes **mutually incompatible** (demographic parity, equalized odds and predictive parity are *almost always* mutually incompatible)
- **Unfairness at individual level** not addressed
- Need to have access to the **ground truth for the target** (for most metrics) and a **representative sample**



Other approaches to measure fairness

- **Individual fairness**: "similar predictions should be made for similar individuals"
- **Minimax Pareto fairness**: "unnecessary harm should be avoided by reducing the risk of the least-favored group instead of making the risks equal for all groups"

Source: Minimax Pareto Fairness: A Multi Objective Perspective, Big Data and Social Science

Bias mitigation techniques (1/2)

Many different bias mitigation techniques are available

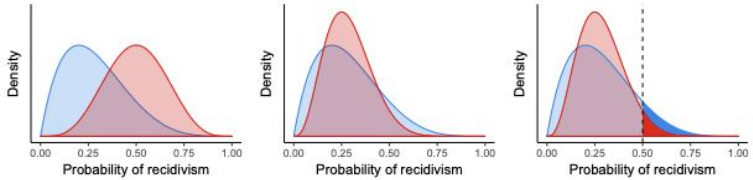
3 main categories of bias mitigation techniques

- **Pre-processing techniques**, applied on the training data
- **In-processing techniques**, applied during model training
- **Post-processing techniques**, applied on the ML model's predictions

In general, bias mitigation techniques reduce the predictive performance of ML models. We may have to find the **right tradeoff between fairness and accuracy**.

Thresholding (post-processing)

Principle: apply a **group-specific threshold** to go from the score provided by the ML model to a binary decision



With a group-specific threshold, enforcing equal opportunity (in the mathematical sense) is straightforward.

Source: [The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning]

Bias mitigation techniques (2/2)

Many different bias mitigation techniques are available

Regularization (in-processing)

Principle: add to the training loss a **regularization term** penalizing the deviation from fairness

Examples:

$$Loss(X,Y) = L_{primary}(f(X),Y) + \lambda Correlation(f(X), A|Y=0)$$

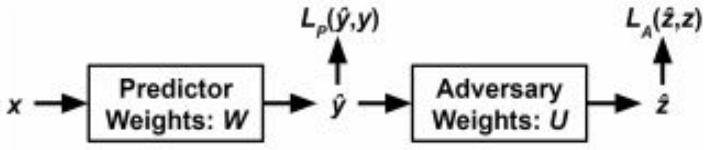
$$Loss(X,Y) = L_{primary}(f(X),Y) + \lambda MMD(f(X_0), f(X_1)|Y=0)$$

… where MMD (Maximum Mean Discrepancy) is a statistical test for the statistical dependence between two sample distributions.

TensorFlow

Adversarial learning (in-processing)

Principle: make an **adversarial network** try to retrieve the sensitive attribute using the model's predictions as inputs (and possibly other data)



Only providing the prediction enforces **demographic parity**. Adding the target's ground truth enforces **equalized odds**.

Source: Toward a better trade-off between performance and fairness with kernel-based distribution matching, Mitigating Unwanted Biases with Adversarial Learning

Limitations

Fairness tools are not silver bullets...

Limited coverage of the ML project lifecycle

The fairness tools assume that **the dataset and the success metric are fixed**.

In practice, **collecting or annotating additional data** as well as **defining the success metric** are important levers to address fairness issues.

Potential mismatch with the legal notion of discrimination

Even if it effectively reduces biases from the perspective of a certain fairness metric, a bias reduction technique could be **considered as a form of discrimination**.

Example: **thresholding**

Only valid with certain assumptions

The fairness tools are **only valid under certain conditions**. E.g.:

- Some bias reduction techniques are **only applicable for certain metrics**
- Computing the fairness metrics requires **reliable data**

Certain ML fairness issues are not captured

The use of fairness metrics would probably not have revealed problems in the following examples:

- **STEM ad targeting** (cf. above)
- **Xerox candidate filtering system**

Source: How algorithms rule our working lives

Key takeaway 5

A growing ML fairness toolbox is available

- The **academic research** community has significantly intensified its effort on ML fairness issues over the last few years.
- Many fairness tools are readily available as **open source packages**.
- **Choosing which notions or which tools to use** in practice is **challenging**. Some of these notions and tools correspond to specific visions of the idea of fairness.

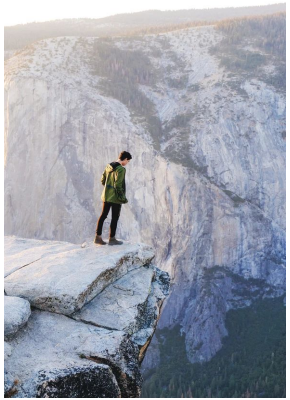
Key takeaway 6

Mathematical tools cover only some fairness issues

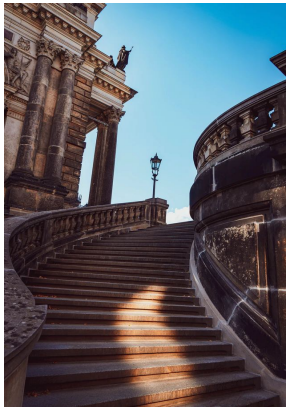
- Mathematical tools to detect and mitigate fairness issues are only valid when certain **assumptions** are met.
- They may **not** be **aligned** with the **legal understanding of "discrimination"**. Using these tools may even be considered as a form of discrimination.
- These tools are **inadequate for certain ML fairness issues** and cover only part of the ML project lifecycle.

Main messages and conclusion

Harms can take many shapes

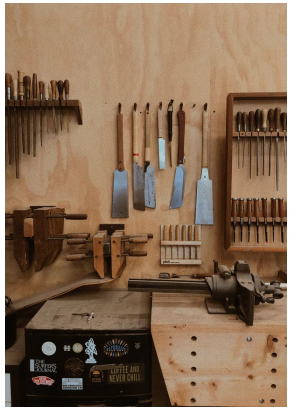


All organizations are vulnerable



Responsible AI concerns the whole ML lifecycle

It's a team sport



A growing ML fairness toolbox is available

Mathematical tools cover only some fairness issues

data iku

THANK YOU