Ethics

• From the Cambridge Dictionary

" the study of what is morally right and wrong, or a set of beliefs about what is morally right and wrong"

• Meta-ethics is about investigating the origins and meaning of ethical principles.

• Applied Ethics is about examining specific controversial issues, such as artificial human reproduction, animal rights, nuclear war, robotic actions, etc.

• Normative ethics is about developing a set of rules that govern human conduct, establishing how things ought to be, and determine right from wrong.

Normative ethics and software/AI

• Highly relevant for understanding and applying ethical principles to the design of software/AI systems • Several schools of thought within normative ethics, among which:

• consequentialism • deontology • virtue ethics

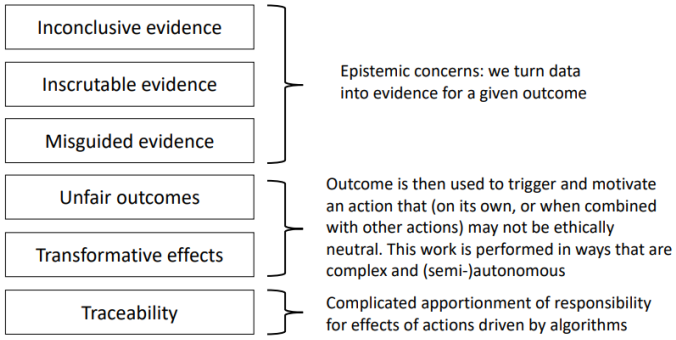
Normative ethics – schools of thought

• Consequentialism (or Teleological Ethics) focuses on the outcomes an action, investigating morality in accordance with the action's results.

• Deontology defines rules for judging the morality of an action, regardless its consequences, and considering duties and the rights.

• Virtue Ethics focuses on the character of a person/group of persons, identifying virtues and wisdom to deal with virtues conflicts.

Possible ethical issues



-Epistemic concerns: we turn data into evidence for a given outcome

• Inconclusive evidence • Inscrutable evidence • Misguided evidence

epistemic concerns that are arise whenever the predictions of use of the system the classifications of uses Software System are either not reliable from a statistical point of view for instance are not inspectable are not scrutable can we have a look for instance to the training data of chat GPT no this is a fact I'm not saying that we should or we should not but given the impacts of this tool that is a data driven tool maybe maybe we should we should start debating this this question but we should able to inspect the training data of system that has a huge impact on the way in which we work and we live or wrong evidence wrong predictions totally wrong predictions the Netherlands prime minister had to resign a couple of years

ago because of a scandal of software automation that was wrong it was asking back to thousands of thousands of families asking back money for welfare assistance of the Gods for their children because the soft automation was doing wrong inferences then it was it was related but this is something different but I had to resign so the automation of software Based on data is reached the level that is quite important for the way we live

-Outcome is then used to trigger and motivate an action that (on its own, or when combined with other actions) may not be ethically neutral. This work is performed in ways that are complex and (semi-) autonomous

• Unfair outcomes • Transformative effects

and then other group of problems ethical problems some Fair outcomes outcomes that are different based on age of people based on the ethnic group they belong to the nationality so the same software is making a disparate impact this is the right term based on personal characteristics so if if I automate the selection of candidates and then analysis of the results of my CV screening tool show that I I discard with a higher frequency women rather than men this is a discrimination that is punishable by law and these are real cases or if I show more often an advertising about a job offer or our house renting offer more often to white people rather to black people this is a discrimination and this is a real cases it will be studied so this is unfair outcomes this is a category of ethical problems and that they can lead also to transformative effects because we change reality if we deny opportunity to People based on their personal characteristics this will have transformative effects on the life of the individuals but also on the evolution of society

-Complicated apportionment of responsibility for effects of actions driven by algorithms

• Traceability

and the last group of problems that will not touch a lot during this part of the course is traceability the who is responsible for what if chat GPT says that I am a thief who is responsible we don't know but it is say something false a guy about myself so we and if chat GPT is used inside a pipeline that gives then it is connected to some certain so some automatic action automatic decision this is a huge problem because I cannot track the China responsibilities this is just an example from a very recent cases in Australia a major was switch it open AI because of false information about himself the GPT said that it was involved in a scandal tax um tax Scandal a professor in US was told by shafted GPT to be involved in a sexual harassment scandal whose responsibility is open questions open problems we are at the very age of research and a social debate

Chapter 02-Data Ethics: premises

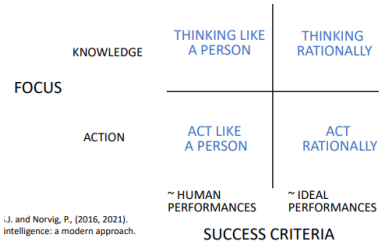
Introduction

The impact of software

• Marc Lowel Andreessen coined two powerful metaphors:

- (2011) Software is eating the world
- (2016) Software is programming the world

Artificial Intelligence



.J. and Norvig, P., (2016, 2021). intelligence: a modern approach.

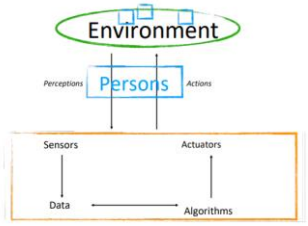
AI as acting rationally: system of "smart agents"

• An agent is something that perceives and acts in an environment. The agent function for an agent specifies the action taken by the agent in response to any percept sequence.

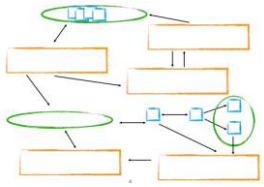
Commented [1]: مشاهده و درک کردن

• The performance measure evaluates the behavior of the agent in an environment.

• A rational agent acts so as to maximize the expected value of the performance measure, given the percept sequence it has seen so far



AI systems as distributed, socio-technical systems



that means that they can exchange information with other systems in different locations and they are called social Technical Systems because people are involved technical system is uh the for instance the telepath systems in Italy in the highway it recognizes your plate you like you play it in your or or of the car it connects to another system to check whether you are loaded to pass or not and then it has an actuator it can allows you grant you the passage or not and then needs to be connected to accounting system

Characteristic of AI agents

• Autonomy: the capacity of an agent to make an autonomous action in an environment in order to meet its design objectives

• Adaptability: the capability of learning from data, and to be able to react to changes in the environment (autonomously, or not).

• Interactivity: the capability of an agent to interact with other agents, be they human or artificial

Automated Decision Making (ADM) systems

• Systems of automated decision-making (ADM) are always a combination of the following social and technological parts:

- A decision-making model
- Algorithms that make this model applicable in the form of software code
- Data sets that are entered into this software, be it for the purpose of training via machine learning or for analysis by the software
- The whole of the political and economic ecosystems that ADM systems are embedded

ADM systems in everyday life

• Predict special life events

• Predict risk of violence at home

• Suggest which university to attend

• Identify the best suitable candidates for a job position

• Detect social welfare frauds

• Take content down from a social network

• Decide about defendants' parole

• Select patients for intensive medical care

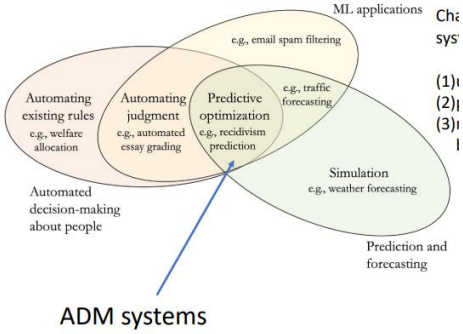
• …

Characteristics of ML/AI based ADM systems:

(1)uses machine learning

(2)predicts future outcomes

(3)makes decisions about individuals based on those predictions.



Expectations on ADM w.r.t human decisions

• High accuracy in predicting outcomes

• Fairness across individuals

• Efficiency gains by reducing time spent by human decision-makers

Data and ADM systems

• ADM systems "learn" from historical series or examples and make (or suggest) decisions. In order for the software system to learn correctly, the following things are necessary:

  • a sufficiently large number of examples

  • a sufficiently heterogeneous set of examples

  • examples annotated with the "right answers"

• Learning by examples is an **inductive** process

Types of inferences

• Deductive All As are Bs. a is an A. Hence, a is a B.

"In deductive inferences, what is inferred is necessarily true if the premises from which it is inferred are true; that is, the truth of the premises guarantees the truth of the conclusion"

this is the way of deduction logical inferences starting from some **facts rules** and then we we deduce we infer some other facts

• Inductive

96 per cent of the Flemish college students speak both Dutch and French. Louise is a Flemish college student. Hence, Louise speaks both Dutch and French.

"Inductive inferences […] are based purely on statistical data, such as observed frequencies of occurrences of a particular feature in a given population"

based on **observed probabilities**

• Abductive

I observed many gray elephants and no nongray ones The best explanation for why I have observed so many gray elephants and no non-gray ones is that all elephants are gray I infer from this that all elephants are gray

"in abduction there is an implicit or explicit appeal to explanatory considerations […] there may also be an appeal to frequencies or statistics. "

• Inductive -> Non necessary Ampliative -> The conclusion goes beyond what is (logically) contained in the premises S

• Abductive -> Non-monotone -> it may be possible to infer certain conclusions from a subset of premises S

02/05/2023

Some problems of inductive inferences…

• When modelling human characteristics and behaviors, there are some potential problems in the use of historical series

• Reality is a super-set of what is measurable

• Survival bias or “low hanging fruit”

that is you measure only what you can measure reality is much much bigger than what is measurable and so this is a non-limitation of such systems that they are based on variables and on data that are a small portion of reality

• Some aspects of our societies (and our life) are measurable only indirectly

The second is highly connected to the first limitation that we cannot measure at all some aspects of our lives and of our society while the first aspect says that reality is so big that you cannot measure all of it . the second aspect says that there are some aspects of reality that are not measurable at all even if you could have all the technical means you cannot measure for instance happiness you cannot measure an opportunity for a person that is located in the future you can make a prediction you can use some proxy but I cannot measure for instance what is the probability that you will become an engineer of success I can make the prediction but I cannot measure it and this is a second important aspect that we see even if it's quite simple let's say actually is present in many systems deployed in reality

• Societies have **historical and structural inequalities**, reflected by the data

جوامع دارای نابرابری های تاریخی و ساختاری هستند که توسط داده ها منعکس می شود

third aspect is that even if we have a very precise measurement system in technological system that can measure several aspects of reality, if reality has some structural inequalities, your measurement will reflect these structural inequalities so as you know computer science is a discipline that is highly polarized with respect to gender I can know all I can measure I can have data about all the individuals in the world that are studying computer science that's fine but the data will reflect the important gender imbalance, so I will have a data that is mostly related two men and four the structural inequality between the global North and the global South the world most of the data will probably come or will be more accurate probably from where from the global North that are economy Society is more advanced from the point of view of Technology deployment.

• Spurious correlations and confounding factors

همبستگی های جعلی و عوامل مخدوش کننده

There is mathematical proof that the bigger and the higher the data sets are the probability to get correlations just by chance are the so-called confounded also confounding factors.
Spurious correlations are correlations that are just random that exist in your data.
confounding factors are factors that you are not taking into account but they do have an impact on your data.

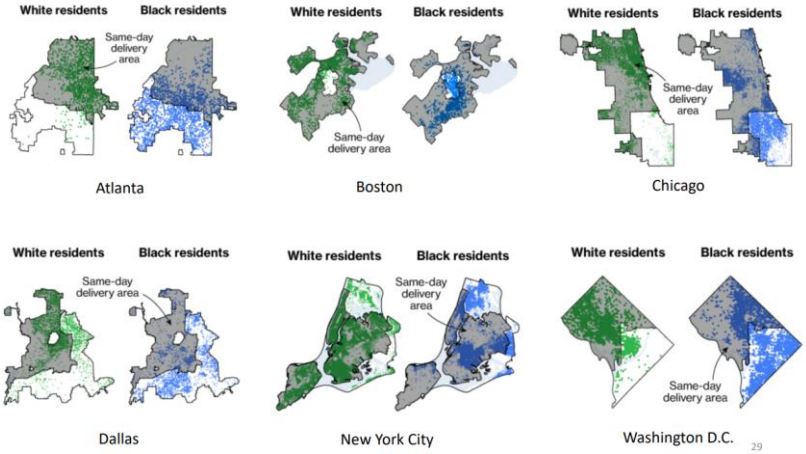
… that ADM systems inherit

An example: demographic disparities

A well-known North American e-commerce company analyzes historical data of online purchases on its platform and demographic data to determine in which neighborhoods to offer the fast delivery service

The first case is from a well-known e-commerce company that is **Amazon.** The company at the time had already vast amounts of information or historical data about purchases of the customers so how about what, when, where, Etc, decided to launch a very novelty and innovative service of the prime Amazon. in the same day the company decided to use to leverage the large amount of historical data that owned about purchases in its platform with the socio-economic data from

the Census that the U.S is publicly available and other economical data not all of this data was possible the time and this article was written to know only the the high level categories of the data that was integrated with the historical data of Amazon the goal was to make a predictive optimization that means make predictions on when or where the probability of people of buying things and requiring goods was very high so that it was commercially profitable to offer the same day delivery that is a the Tesla cost because you need to have a lot of transportation means in order to offer this possibility.



What was the result?
The result was an unpredictable impact so that the first city was Atlanta and for each City you see two maps, one that is filled with green dots and the other one is filled with blue dots. the one that is filled with **green dots** represent a unit of population of **white residents** so they belong to the category white calculation that is an official category used in the U.S census (سرشماری), the **blue dot** instead is referred to unit of population of the same size about **black residents** that means the people that in the U.S census is labeled as Afro-American people.
the **gray area** is where the same day delivery is offered the white area or **non-filled area** is where the service is not offered you can see in Atlanta most of the people that were taking advantage of the same day delivery well white people because you can see in the gray area just focusing for a moment on the gray area if you check the density of the green dots is much much higher the not only the density but also the the number of green dots is much much higher than in the black residence Maps where blue dots are less also if you take into account the totality of the green dots roughly more than 95 percent are all located in the gray area that is the other where the services offered if you take into account only the blue dots approxes approximately three over four 75 percent of the blue dots are out of the area where the service was offered this was an implication on a unplanned implication and impact of the statistical optimization algorithm that decided that this was the most profitable area to offer the service and in this area most of the white residents were living.

We can debate whether a company like Amazon should worry about these aspects and this is a fact. this is an impact of the disparity, impact of the algorithm, the decision on where to offer the service or not was impacting differently, people according to a certain attribute which is a special attributes in the skin of the color it's called a protected attribute
because you have to guarantee some Equity towards ethnic origin and other attributes as you will see during the the course for us for a certain type of services this service is not among those services that should take care of protected attributes so let's just make it this clear in any case the impact was notable and was covered by this journalistic investigation we can check quickly all the other five cities in **Boston** in the small area which is quite a rich city so in most of the city the service is offered unless these smaller Central area where mostly blue dots are present so again the white residents they add all access to the service to the fast delivery service in **Chicago** the situation is quite evident that in the white area very very few green dots are available why are represented while in the corresponding black residence map quite a remarkable number of people with dark skin or could not use this service same for **Dallas** same for **New York City** although less clear than other cities same for **Washington DC** but this is an area where most of the black people lives the algorithm was simply optimizing by the profitability so the area where residing most of the people who could spend the money on the platform

In many cities, white residents were twice as likely as black residents to live in a neighborhood where service was offered.

After the publication of the study, the company extended the service to many of the districts that did not have it, in the cities mentioned by the study

A few considerations

• The classifications of the system had a disparate impact on black people

• Neither the source code nor the design of the system are inspectable

• The data used by the software is also not publicly available

Another example: the COMPAS case ➔ very important case

The **COMPAS*** system is used in some U.S. court systems **to estimate the probability of recidivism of a convicted offender.**

*Correctional Offender Management Profiling for Alternative Sanctions

This is a very important case that inspired hundreds of scientific studies in the machine learning communities Compas.It is an acronymous **Correctional offender management profiling for alternative sanctions,** what does it mean?
it means that it's a software that recommends judges about whether to give an alternative sanction to a person who is charged for a crime so instead of staying to prison I can have a limited Freedom limited field or I can pay we are in the justice system and terms are very technical, so for paying to Be Free as a specific technical term in the Justice area.
However, the idea is the following one to support judges and make judgments more fair so recommending them what a risk indicator for each person.
So imagine that you are a judge and you need to decide about a case in front of you. you have a person and you need to decide whether this person that was stopped for either a crime or a suspected crime should stay in prison or can have a certain level of alternative sanctions such a level of freedom the score that the person can receive is from 1 to 10 and 10 being the highest possible it means high risk of what a risk of reoffend in the future so the system makes the probability based prediction or whether a certain person will commit a CR will make a crime again in the near future if yes, if the probability is I then is the risk score will be 10 9 8 high risk otherwise we make we will output the lower risk scores from one to four lower score five six seven medium risk. This is the idea of how the system works. It's not substituting judges, it is helping them, it is supporting them in this decision. This is an example really.

The (main) problem in numbers

	White	African American
Labeled Higher Risk, But Didn't Re-Offend	24%	45%
Labeled Lower Risk, Yet Did Re-Offend	48%	28%

the 24% of white people were liable at a higher risk but did not reoffend these are the **false positive**. so I receive a high risk but then I will not commit a crime in the next two years. this is not so this is not a major of how much you are dangerous this is just the prediction of whether you will enter again the criminal record.

45% **False Negative** African-American

Same considerations as before

• The classifications of the system had a disparate impact on black people

• Neither the source code nor the design of the system is inspectable

• The data used by the software is also not publicly available

Given the circumstances…

• …which other information should be displayed to judges ?

• …what alternative classification goals are more suitable for the judiciary context ?

• … which stakeholders should be included in the correction of the system ?

• … who is responsible for software misclassifications ?

Other examples

* image that's a was quite become quite viral on Twitter a few years ago about a **soft dispenser that did not recognize the hand of black people** probably because it was designed in a certain way to recognize only light skin this is a disparent impact.

* another example of this impact is an example of searching of images that are given by Google when you make a **search with the word CEO** only made white here the impact is less evident and there is not a direct impact on decision attached to this software output however there is a certain representation of reality it is quite polarized no women here well here one.

* another Google Images search with the word **Carpenter nurse** so Carpenter only male nurse only female in English this term is **neutral** and is not connotated by gender.

* this is a tool software classification image classification by Google data tells you what is in the picture so here there is a hand and a gun but if we change the color of the hand and we make it clear there is a hand and a monocular not a gun anymore another example of this product impact

very few terms one just one in some cases the level of **toxicity changes** quite a lot

* if you ask her for an image of an **emotional person** you will get basically images from **women**

*if you ask her up an image of a **poor person** you have a quite clear

*indication of the ethnic group groups attack they are all **black a terrorist** s they are well connotated this is output from Dal e

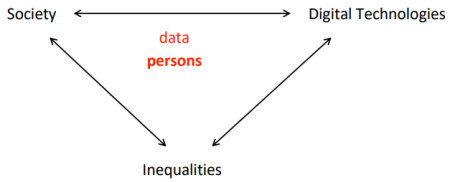
"[…]trained on biased data then they can learn to repeat those same biases. Sadly, our societies have such a history of prejudice that you need to work very hard to get that bias out"

the problem is an acceptance from a public speaker of Margaret Vistager who is in charge of competition law and digital in Europe and is also the vice president of the European Commission. she says that's very clear what is the problem trying on biased data these kind of systems automated decision-making systems.

they can learn to repeat those same biases and sadly our societies are such a history of prejudice that you need to work very hard to get that bias out I don't know whether it is possible or not to remove this bias, because if you change the training data you lose accuracy so it's a problem it's an open problem for businesses and research as well but the problem is now known at the highest possible level of the European institutions

او می گوید که بسیار واضح است که مشکل تلاش بر روی داده های جانبدارانه این نوع سیستم های سیستم های تصمیم گیری خودکار چیست . آنها می توانند یاد بگیرند که همان تعصبات را تکرار کنند و متأسفانه جوامع ما آنقدر تاریخ تعصب دارند که برای از بین بردن آن تعصب باید سخت تلاش کنید. داده‌های آموزشی دقت خود را از

دست می‌دهند، بنابراین این یک مشکل است، یک مشکل باز برای مشاغل و تحقیقات نیز هست، اما این مشکل اکنون در بالاترین سطح ممکن از موسسات اروپایی .شناخته شده است



relationship between **society** and **digital Technologies** a specific type of digital technologies that we defined in terms of **inequality**. our question is to understand whether the predicted optimization systems that have been deployed a lot in these years are increasing or not are improving the inequalities of our society? unfortunately I have to anticipate you that the question is the answer is negative they are not improving they **are making things worse** and we'll try to understand why we're trying to to have a Notions of agility fairness to understand how to measure the fairness of all algorithms and why because our discipline is quite new

Complex socio-technical systems

EU Charter of Fundamental Rights Article 21 - Non-discrimination

1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of **nationality** shall be prohibited

take this list of attributes as a reference point whenever we speak about discrimination whenever you you will see in a exam text explain possible discrimination issues they should all refer to protected attributes not to any type of attributes in a database so this is a very important article it's a particle comma 1 of article 21 the European Charter of fundamental rights

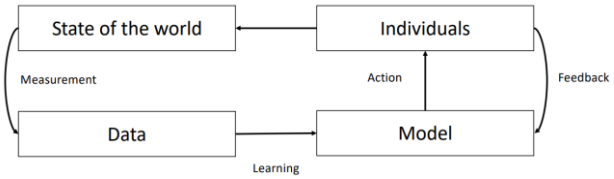
for instance in a country where it is difficult to have a car because it's very expensive when building such a huge highway or even streets Urban Street you need to take care of people that cannot afford a car so probably sidewalks should be a bit better than these to avoid having a negative strong impact on the life of people.
so this is a metaphor but in the discipline of urban constructions, there are techniques and awareness of this kind of social issue and this kind of impact especially towards vulnerable people like any other engineering discipline.

Commented [ZK2]: خیلی مهم : حتما در امتحان در توضیحات به این لیست و قانون اشاره بشه

1. هر گونه تبعیض بر اساس نژاد جنسیت ،تولد ویژگی های ژنتیکی، زبان و مذهب و... ممنوع است.
2. در محدوده اجرای معاهدات و بدون لطمه به هیچ یک از مقررات خاص آنها،هرگونه تبعیض بر اساس تابعیت ممنوع است

Chapter 03- Demographic disparities in the loop

The machine learning loop



first of all the **individuals**, ask the people together, then make the so-called - **states of the world -** the state of society where we can aggregate data on individuals and elaborate some patterns on groups of individuals so this is the reason why they divided the two elements by means of **measurement**.

you **get data from the state of the world** and so **from individuals**, you **learn** from **data** and you **build the model** as an output, a **prediction or classification** that is bound to **action**. This is very similar to the definition of artificial intelligence agent that I gave to you so there is an action on individuals.

this Clause is the big loop then there is a shorter Loop because individuals can change their behavior according to the action of the model . the system that implements the model and they can provide **feedback** that updates the model directly.

this is possible technically not easy but possible this creates a further Loop inner loop this is a formalization.

Commented [ZK3]: ما حالت‌های به اصطلاح جهان را حالتی از جامعه می‌کنیم که بتوانیم داده‌های افراد را جمع آوری کنیم و الگوهایی را روی گروه‌هایی از افراد شرح دهیم، بنابراین به همین دلیل است که آنها این دو عنصر را با اندازه‌گیری تقسیم می‌کنند.

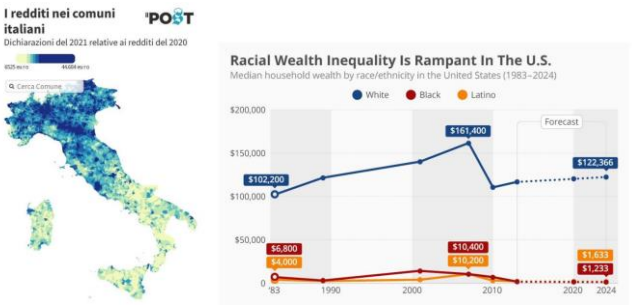
شما داده‌ها را از وضعیت جهان و بنابراین از افراد دریافت می‌کنید، از داده‌ها یاد گرفته‌اید و مدل را به‌عنوان یک خروجی پیش‌بینی یا طبقه‌بندی می‌سازید که مفید به عمل است.
این بسیار شبیه به تعریف عامل هوش مصنوعی است که من برای شما ارائه کردم، بنابراین یک اقدام در مورد افراد وجود دارد

Demographic disparities

- Disproportions and inequalities are common in our societies. E.g.:
    - Certain groups of individuals may concentrate in specific neighborhoods of a city, or of a geographical region
    - Certain groups of people use digital technologies more easily and frequently than others
    - Services to citizens and industries are unevenly distributed

As we have already said, society is not it's not perfect so there are some **structural inequalities.** For instance in Italy it is well known that the economy is much much stronger in the north than in the south and if you collect economic data on individuals this will be most of the time reflected.

- Sometimes they are not inequalities, just **disproportions** like the number of women in this class is probably 10-15 percent because this is common in the population of computer science courses.
- we have seen that in the **Amazon case** that the probably the the black people were living in and they were all concentrated in areas that were poorer that's why the those area were not selected by the algorithms to offer the same day delivery service



income inequality in italy and USA

بنابراین حدس می‌زنم نابرابری‌های ساختاری را که با هم دنبال می‌کنیم، مبنای استدلال شما در امتحان خواهند بود، من همیشه به نابرابری‌هایی که امروز خواهید دید و روشی که در موارد دیگر خواهیم دید اشاره خواهم کرد.

for the sake of the exam it's enough that you know that there is a huge disparity between white people and other ethnic groups especially black people and Latinos in the United States for instance just to be concrete and practical.

At least half of the UK's black staff are affected by racial pay gap, new research finds, so black people are paid less uh also with the same work duties.

The same is happening in Italy with women with the same duties paid less.

there are also relevant differences in the 12th year and also between other comparisons. so the graph says that education skills are more advanced, setting white people in the U.S. In black people there is a social-technical, social economical explanation and is based on the fact that the school system and U.S are also based on the Economic Opportunity of your family of a region but **this is not a class of inequality, this is a class of data ethics** so what is important for you to know is that **there is a gap of Education skills** U.S and this is a fact between white and the other ethnic groups, especially black people.

we have a problem in the equality in terms of **payment gender equality** in terms of payments

Percentage of gender salary gap in Italy from 2015 to 2020, by grading

Gender pay gap in Italy 2015-2020, by grading

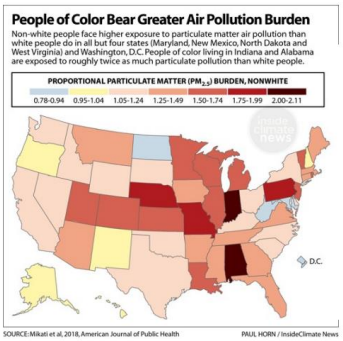
	2015	2016	2017	2018	2019	2020
Top Management	11.9%	12.2%	9.6%	8.2%	8.8%	7.6%
Middle Management	5%	4.4%	4%	3.8%	4.4%	4.5%
White Collars	12.4%	11.7%	9.4%	9.5%	11.1%	11.4%
Blue Collars	11.5%	12.9%	10.8%	10.6%	11.3%	13.3%

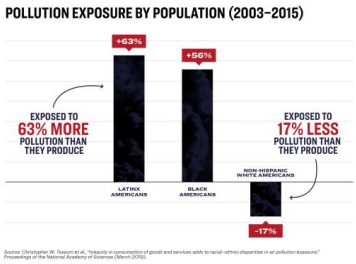
this is a statistic from 2015 to 2020 in Italy gender per Gap women and on average 27 thousand Euros per year and many Thirty thousand

average so to be taken carefully this is average but there is a pay Gap of about three-four percent according to this data that is consistent with the five percent by OCD

if you want to go into a bit more in detail on this data we see that most of the difference is the in the low lower paid jobs blue colors where the difference reaches even at 13 percent so women blue colorless women are paid more than 10 percent less than men at a part of a category of work the difference shrinks until a minimum in the middle management and again increase again in the top management.

I will never ask you about the numbers but it's important that you know about this fact





if you have a less economic opportunities you live also in territories that are more exposed in terms of environmental damages

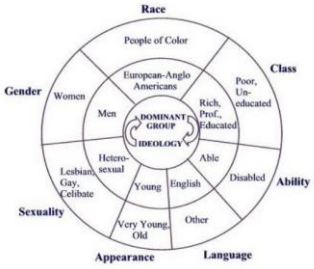
this is a fact that economic opportunities can give you obviously better Medical Treatments, better knowledge on how to take care of yourself, and more money to buy medicament.

there might be a correlation between your socio-economical position and the level of Medical Treatments that you can afford. especially in countries where there is no Public Health System and you need insurance but if you don't have work you don't have insurance.

medical insurance is a big problem U.S again because people who cannot afford medical insurance cannot succeed a certain medicines

The matrix of domination/oppression

- The sociologist Patricia Hill Collins refers to race, class, and gender asan interlocking systems of domination and oppression
  - Collins, P. H. (1990). Black feminist thought in the matrix of domination. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment*, *138*(1990), 221-238.

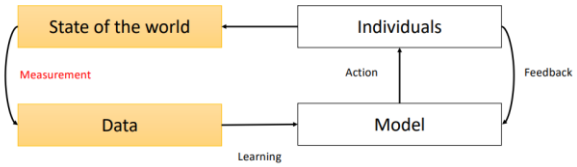


The matrix of domination explains in general in the most advanced economies

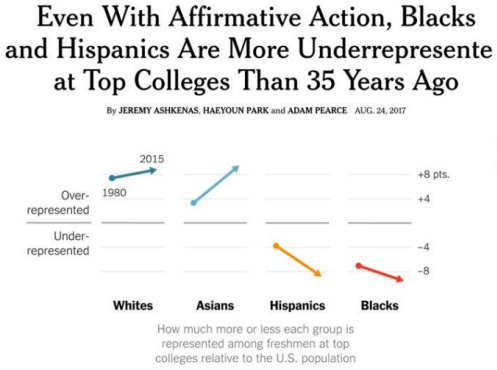
what are the characteristics of the dominating classes? What is the dominating class? it is the social class that owns most of the political and economic power and the dominant group.

it is important to understand that wealth is racially connotated, income is racially connotated. in many countries can be even a gender connoted as we have seen so this kind of structural inequalities should be taken into account when using predictive optimization for social outcomes because social outcomes are the results all the structure of the society in which you live

State of the world → Data



it should be clear now that the world has some demographic disparities and some structural inequality States can be transferred from measurement into Data into the training data of the data driven system


Even With Affirmative Action, Blacks and Hispanics Are More Underrepresente at Top Colleges Than 35 Years Ago

By JEREMY ASHKENAS, HAEYOUN PARK and ADAM PEARCE AUG. 24, 2017

How much more or less each group is represented among freshmen at top colleges relative to the U.S. population

US Universities started using the "multiracial" category only in 2008

Race as a social construct

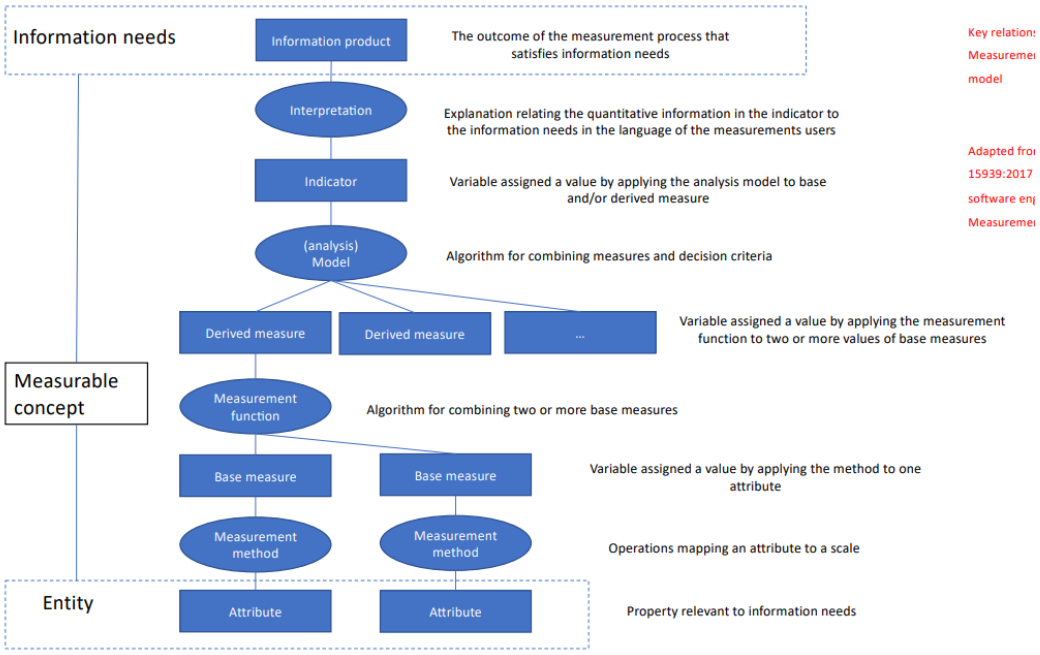
• Racial classification in South Africa as demanded by legislation passed in 1950 (and implemented by IBM):

- Europeans
- Asiatics
- Persons of mixed race or coloured
- Natives or pure-blooded individuals of the Bantu race

The measurement process

• Data is often given the characteristic of "objectivity". However, data about society and individuals is the result of a measurement process, in which many subjective choices have to be made.

• The measurement process is the empirical process of assigning (numerical) values to an entity, with the purpose of characterizing a specific attribute.



first of all the measurement process is an empirical process which you bind an information quantity to an entity so you have an information should be measurable

Entity is **people, persons, or groups** made by attributes. they need some **information** about them so I take some **attributes** and each attribute needs a proper **measurement method** to derive a **base measure**.(e.g:criminal record, number of prior crimes, unemployment status)

Measurable Concept
To evaluate the recidivism risk of a person, I need at least criminal history as the number of crimes in the past and the unemployment status. I need to separate **measurement methods** to track the number of previous crimes.
I should also just Define where I found this information and what type of primes I needed to track. This is the very first decision in which identifying an attribute of an entity even identifying attributes that are relevant for our purposes are **subjective choices**.

base measures can be taken directly as they are or it or can they can be combined by a measurement function in order to get a derived measure. I may take the number of Prior crimes as they are 10 20 15 or I might decide that the absolute value is not fair and I want to normalize it by the age of the person and this is an example of **derived measure** and this is of course we'll have an impact at the end the rational of normalizing the number of crimes by the age is to have uh the same to say this the same starting level for everybody because a person of 55 years old might have more probability to done more crimes than a person of 18 years old because of Simply of the age and this is an example of a very simple measurement function that I apply to a base measure but here again this is a very subjective choice that needs to be explained needs to be documented I can have a set then of the right measures and some of the derived measures can be directly developed **base measures** and they are combined into a model analysis model or predictive model or a classification model that should output number an indicator risk of racism number from one to ten again this is a here we might have the choice to decide the scale or not depending on what type of algorithm a learning algorithm I have here the indicator might be taken as it is or there might be a further interpretation that's a so that gives the another information product so the outcome of the measurement process that satisfies what were the initial information needs and the example of the relative risk we have seen that although the risk the output of the algorithm is from 1 to 10 then it is bound to a further scale **low medium and high risk** and this is an interpretation of the value.

as you see now there is this very simple schema gives us all the Analytical in a way it is analytical all the steps in which a human decision is involved and often very often all these steps are either implicitly done or not documented at all so usually when you get a data set you don't have all the documentation of all these choices very rarely that you can have a flow and clear explanation of all these types that were done for every variable every feature and if there is time I will show you a few uh example of this documentation lock and this is this is a problem of of trust of accountability of the algorithm depending on on the case I think on the purpose of the classification on the purpose of the prediction on the

Example

• The management of a company decides to adopt an automatic system for identifying the 10 most productive developers in the last year, and reward them. The following choices are made:
• The company code repository is taken into consideration

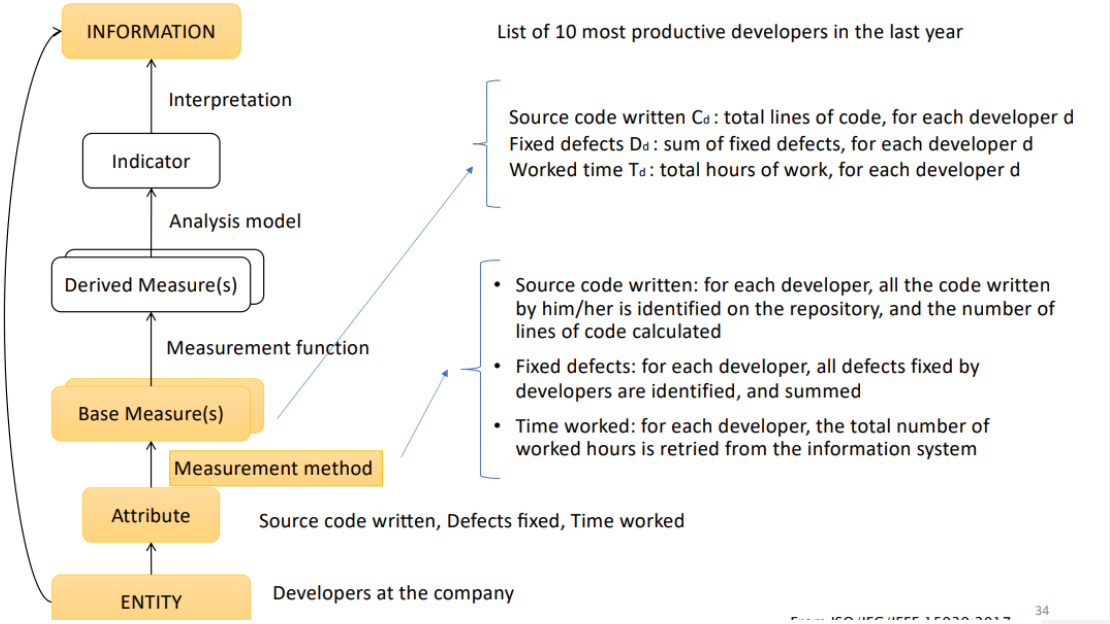
• A second source of data will be used: information on the daily presence of staff in the company
• Productivity must be measured both in terms of source code committed and in terms of fixed defects
• The final choice is made on a unique indicator of productivity The company defines the following measurement process

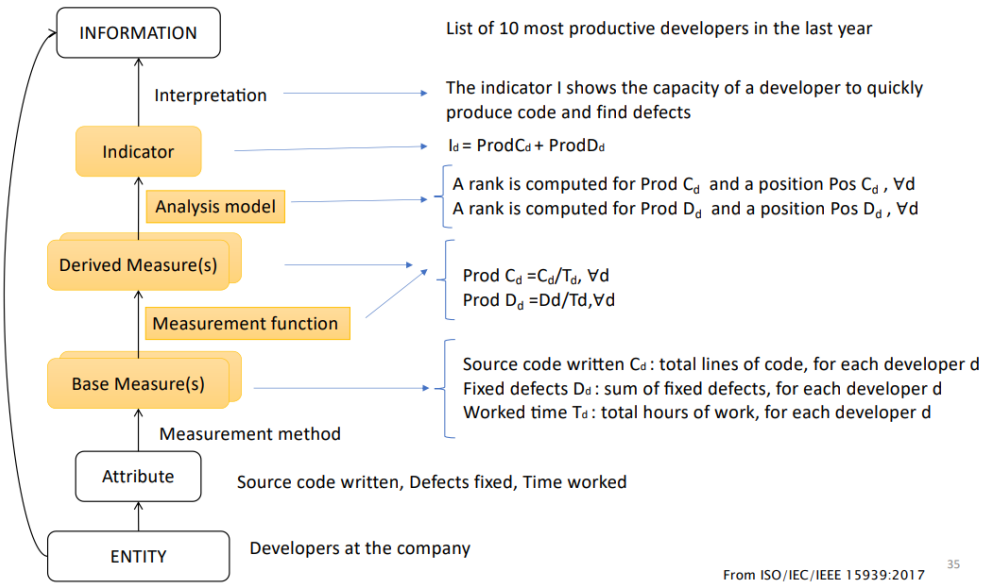
1) **from the information that is needed → List of 10 most productive developers in the last year**
2) **who are the entities?--> the developers of the company**
3) **What articles are chosen? → the source code written, the defects we fixed, and the time worked**

Cd is the product Cd is the total number of lines written divided by the hours spent hours worked for each developer

the next step is to derive the measures to get the right measures of through measurement functions that combines the base measures
the derived measures are a product of the following two measurement functions

Answer to example question





Some considerations

• Some programming languages are more "verbose" than others (if you write a function in Java it will be much shorter that in if you write a function in assembly)

• The indicator does not take into account the]me spent on other ac]vi]es: documenta]on, understanding of requirements, etc.

• Moving from a ra]onal to an ordinal scale eliminates distances between posi]ons

• The more any quan]ta]ve indicator is used for decision-making about people, the more it will be subject to the pressures of corrup]on and the be^er suited to distort the social process it intends to monitor (Campbell's Law, or reflexivity problem)

• Goodhart's law (rephrased by Marilyn Strathern): "When a measure becomes a target, it ceases to be a good measure".

Follow up

• Let us now assume that the following year the management of the company decides to reward groups of developers rather than individually. The average productivity is used, and two managers are commissioned to perform the calculation. The two managers use two different measurement methods for fixed defects:

• the first one uses defects fixed/hour; the second one uses the reciprocal (hours worked to fix a defect).

• The data collected are shown below.

	GROUP 1		
MANAGER	Developer 1	Developer 2	Average
M1 (fixed defects per hour)	1	4	2.5
M2 (hours for a defect fix)	1	0.25	0.625

	GROUP 2		
MANAGER	Developer 1	Developer 2	Average
M1 (fixed defects per hour)	2	2	2
M2 (hours for a defect fix)	0.5	0.5	0.5

Manager 1 awards group 1 because it fixes on average more defects per hour (2.5 vs 2). Manager 2 awards group 2 because on average it is quicker to fix defects (0.5 vs 0.625) Both managers used the same data !

Simpson's paradox

According to Simpson's paradox, a trend, association, or characteristic observed in underlying subgroups may be quite different from association or characteristic observed when these subgroups are aggregated. In the example, the specific cause lies in the non-linearity of the transformation to calculate the reciprocal: if a function f is not linear, then E(f (x)) != f (E(x))

Further comments

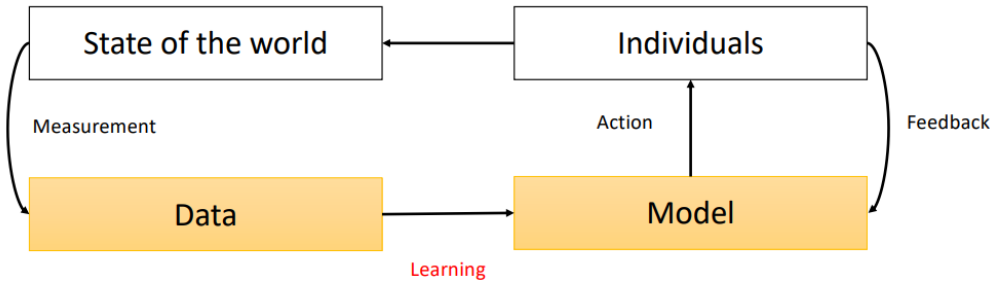
• Measuring involves defining the variables of interest, defining the process to interact with the real world and transforming the observations in numbers

• Usually software developers do not follow the whole process, but they use data according to some given requirements

• Data quality might be not suitable for the intended use (e.g. incomplete data, or wrong, outdated data, etc.).

• Remember: the real world is much more complex than variable-value pairs, and the rules defined to analyze them

Data → Model



Models can propagate demographic disparities in the data

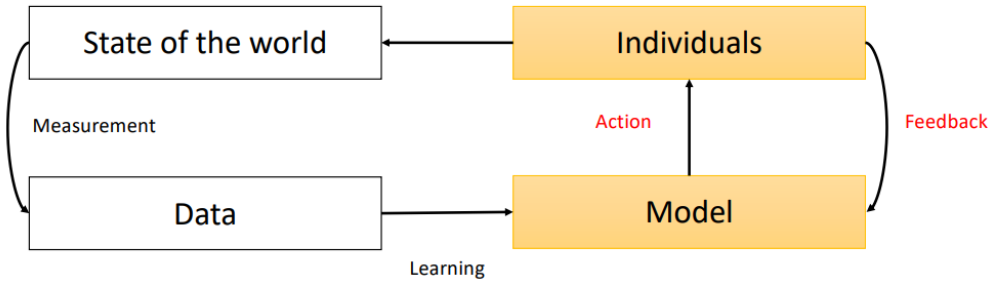
measurement process can either replicate some demographic disparities or inject the father ones or inject some other problem measurement issues. the last step, it is the **model learns from the data or suffer from the limitations of the data sets itself**

«اگر مدل‌های پذیرش در دانشگاه‌های آمریکا بر اساس داده‌های دهه 1960 آموزش داده می‌شدند، احتمالا اکنون تعداد کمی از زنان ثبت‌نام می‌کردند، زیرا مدل‌ها برای شناسایی مردان سفیدپوست موفق آموزش دیده بودند».

imbalanced learning

if you write in English she's a doctor and he's a nurse and you translate it into a language in which there is an instead different pronoun for each gender. then this is the reverse sorry in the English case you have different problems for different genders you translate it into Turkish and you get the natural part and I don't know how to read it then if you translate it back you can try to make this game into Google then you can see that you've translates back into English you get he's a doctor she's a nurse and this is because Google translate simply is trained with data that is imbalanced by a profession this is normal

Model→ Individuals



when our model acts on reality and then when individuals can change their behavior and this change is somehow captured by the model there is an inner loop here

Predictive policing - Preliminary considerations

• The characteristics and behavior of individuals change over time

We are trying to predict something that can change during the time criminality of a place or a person can change, so probably we need this feedback back to the model because otherwise the model will be trained like the model for universities.

• Only certain crimes can be easily mapped

• The output of an algorithm can have an effect on the individual, which tends to confirm or contrast the action (Remember Campbell's Law and Goodhart's law)

Feedback loops

• Feedback loops are also a well-known problem in social media: e.g. more "like" I will use for some categories of news, the more the algorithm will tell me about similar stories, creating the so-called "echo chambers."

digital platforms to the so-called Echo Chambers when in Facebook or in Twitter

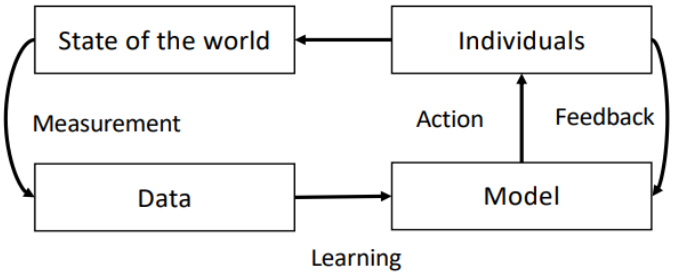
Chapter 04-BIAS-SOURCES

Bias in software systems Computer systems that systematically and unfairly discriminate against certain individuals or groups of individuals in favor of others [by denying] an opportunity for a good or [assigning] an undesirable outcome to an individual or groups of individuals on grounds that are unreasonable or inappropriate .

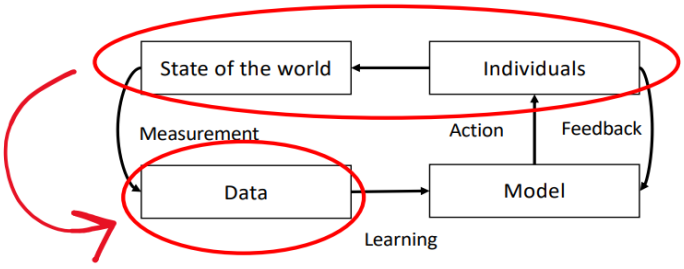
Sources of bias in a typical ML/AI process

A unified framework based on the scientific literature

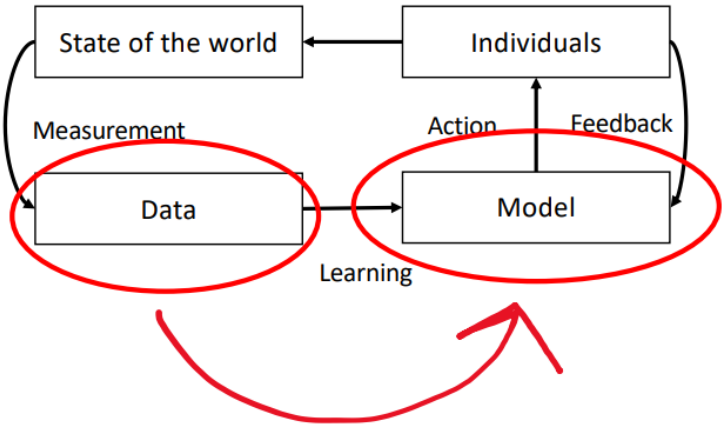
Bias and ML cycle



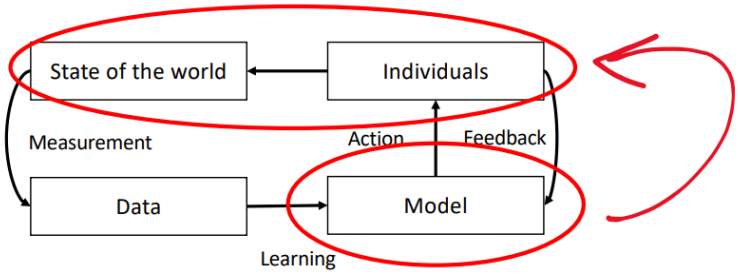
Bias from individuals/society to data



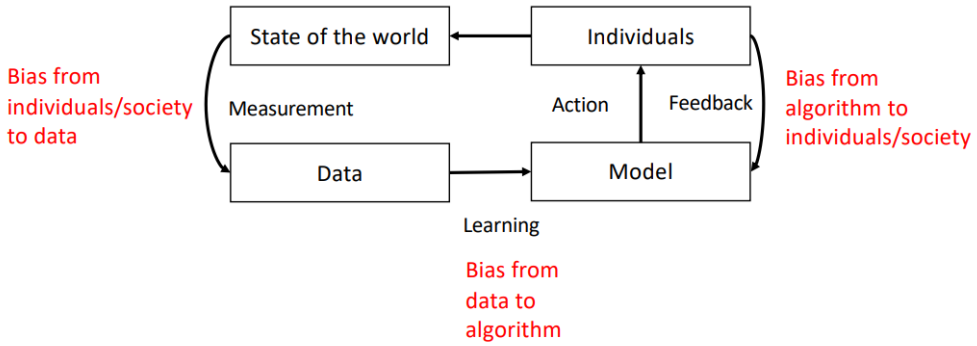
Bias from data to algorithm



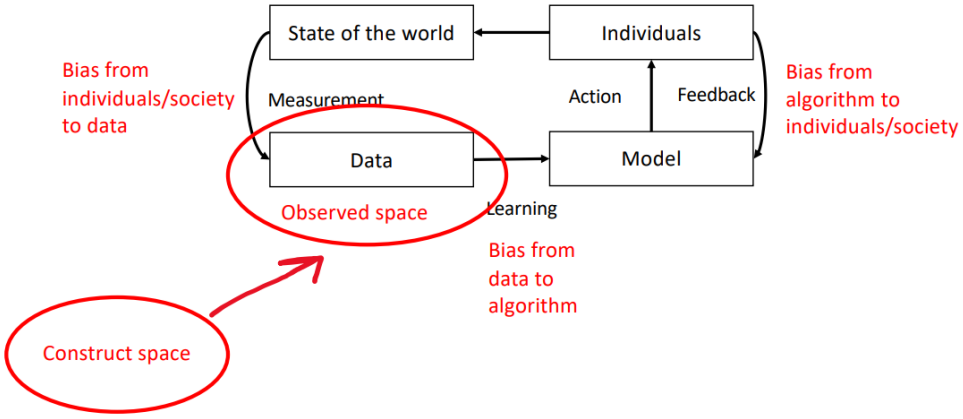
Bias from algorithm to individuals/society



ML cycle and bias



From the ML cycle to the spaces

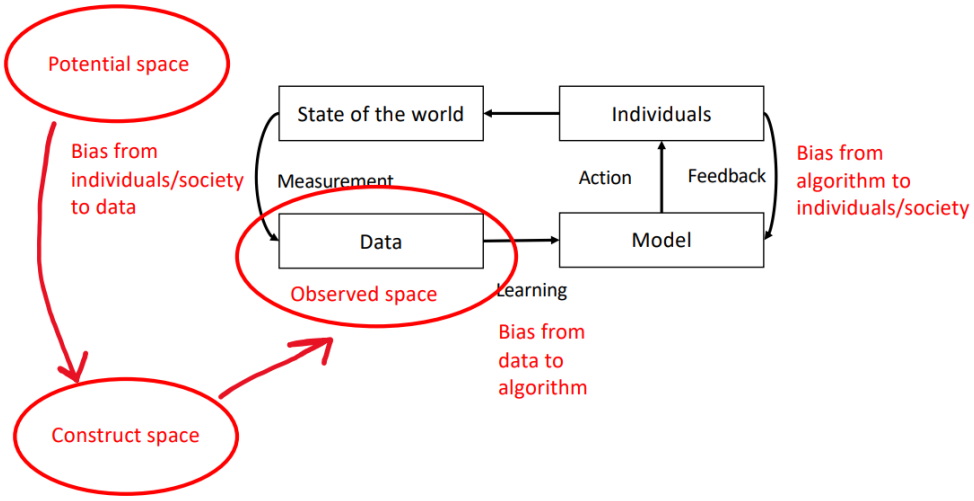


Commented [4]: observed space: what we can observe and comes from what is called the construct space.

Construct space : Its quantification goes into the observed space. also the space where we study the phenomenon.

this is important. So first of all, we have a difference between the constant space and the observed space. That is the space where data is

From the ML cycle to the spaces

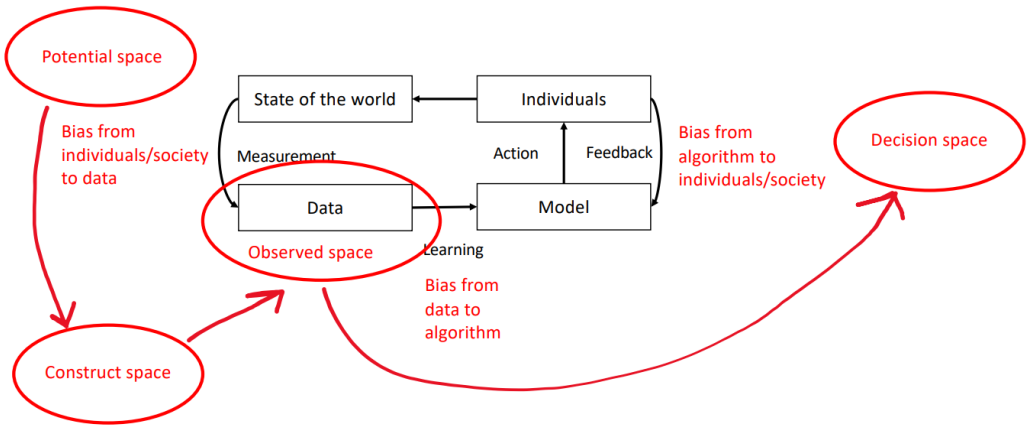


Commented [5]: potential space : If I were African-American, for instance. If I were in another social context. It's called the potential space.
من فکر میکنم منظوراین باشه که بسته به اینکه اطلاعات جمع آوری شده ما از کجا و چه جامعه ای باشه داده ها و اندازه گیری ما متفاوت میشه

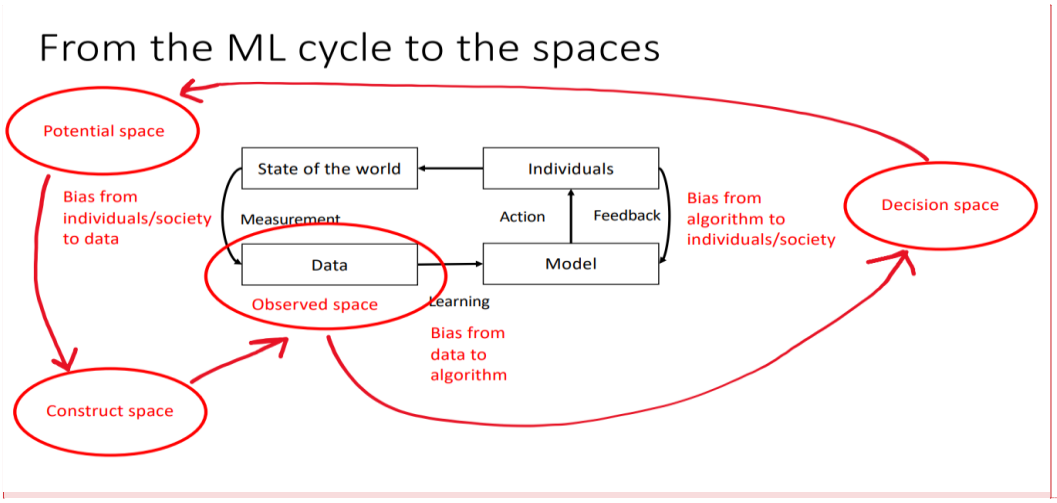
decision spaces: where decision actually occur from the observed space is where the data is used to make some classification predictions that in turn come to be decisions that have impact on the people

going back both to the constant space and to the potential space. We use this further abstraction built on top of the ML cycle to classify the different sources of data.

From the ML cycle to the spaces



Commented [6]: from the potential space to the construct space, we see that the historical bias will occur. So structural inequalities in society that makes potentiality, potential capabilities of individuals less effective because of a variety of economical, societal, et cetera, cultural explanations



Commented [7]: When we go from the construct space to the observe space, we have a lot of subjective decisions that can insert limitations, errors, biases

The data is called the **observed space** so it is what we can observe and comes from it is called the **construct space**, What are the theoretical constructs productivity is a construct is a term that is in the constant space. it's a concept its quantification goes into the observed space.

this is important so we have first of all we have a difference between the constant space and observed space that is the space where data is .even have a potential space how would be the construct space. if I were Afro-American for instance, if I were in another social context. it's called the **potential space.**

and then we have the **decision spaces** where decisions actually occur. The **observed space** is where the data is used to make some classification predictions that in turns can to be decisions that have an impact on the people. so going back both to the construct space and to the potentials space. I will use these are further abstraction built on top of the ml cycle to classify the different sources of data after having done that we will move to a statistical formulation of algorithmic fairness in order to be able to quantify the impacts of the disparity impact of an algorithm.

Formalization

- Y : target variable, the quantity to be predicted
- X: features variables that determine Y
- Phenomenon of interest modeled by the relationship Y=f(x)+e

Space of observations

- Construct space: $Y = f(X) + \varepsilon$
- Data collection→ Space of observations:
  - $\tilde{X} = g(X)$
  - $\tilde{Y} = h(Y)$
- Data from observations: $(\tilde{x}_1, \tilde{y}_1)$, $(\tilde{x}_2, \tilde{y}_2)$, $(\tilde{x}_n, \tilde{y}_n)$
  - Each pair is
    - An instance $\tilde{x}_i$(a vector)
    - A label $\tilde{y}_i$ (in most cases, we will assume $\tilde{y}_i \in \{0,1\}$
- A feature that is a protected attribute is named A

Protected attributes **We will consider those from Art. 21 Article 21 - Non-discrimination EU Charter of Fundamental Rights**

1. Any discrimination based on any ground such as sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation shall be prohibited.

2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited.

Commented [8]: مهم برای امتحان
Remember our definition of protected attribute that we decided to stick to the Article 21 of the European Charter of Fundamental Rights. can be gender, can be race, can be color, ethnic, or social origin, and so on. Any time that I ask for in an exam for a discrimination risk or explanation of a given discrimination, the discrimination should always be with respect to one or more of these attributes. and not on some other attributes that are present, variables that are present in the text.

Commented [9]: هرگونه تبعیض بر اساس هر دلیلی مانند جنسیت، نژاد، رنگ، منشاء قومی یا اجتماعی، ویژگی های ژنتیکی، زبان، مذهب یا عقیده، عقاید سیاسی یا هر عقیده دیگری، عضویت در اقلیت ملی، دارایی، تولد، معلولیت، سن یا گرایش جنسی. ممنوع خواهد شد

Commented [10]: هرگونه تبعیض بر اساس تابعیت ممنوع است

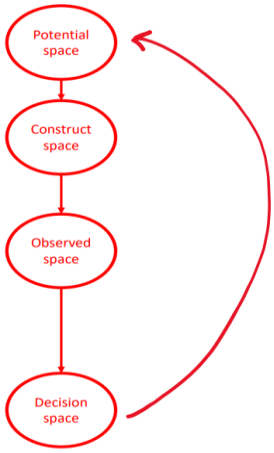
Predictions and decisions

- $\widehat{f} \sim f$ estimate of f
- $R = \hat{f}(\tilde{X})$ prediction R made by the classifier trained on $\tilde{X}$
- $D = d(R)$ decision D made with a decision rule d applied on R
- In certain applications, it might be decided that $D = R$
- or that a decision relies upon:
  - additional environmental information Z
  - considerations about a protected attribute A (or more)
- → $D = d(R, A, Z)$

Commented [11]: این بخش مربوط به decision space است
some specific attributes of the person (A) or some other variables, environment variables, depending on the decision process.

Mapping to spaces

- $Y = f(X) + \varepsilon$
  - Construct space
- $\tilde{X} = g(X)$, $\tilde{Y} = h(Y)$
  - Observed space
- $R = \hat{f}(\tilde{X})$
  - prediction R made by the classifier trained on $\tilde{X}$
- $D = d(R, A, Z)$
  - decision D made with a decision rule d
  - d applied on R, prot. attr. A, env. information Z



Graphical notation: variables

- R: classification
- A: sensitive characteristics/protected attribute
- Y : target
- C: capacity of an individual
  - Example: economic resources, properties, personal talents, skills, etc.

- PC : proxy variable we have access instead of C (or any other variable)
  - E.g., university final grade is a proxy for skills
- Q: additional variables
  - they may or may not be relevant for the problem (i.e. impacting Y)
  - they may or may not be impacted either by R or A,
  - e.g. the neighbourhood where one lives in.

Commented [12]: we can have a proxy variable that you might be able to track instead of the capacity. The quotient of intelligence is a typical proxy variable. We cannot measure intelligence. We use the so-called intelligence quotient, EQ. You can have proxy also for any other types of variables.

Commented [13]: so for the target variable or for the protected attribute data or for any additional variable that we will call Q.

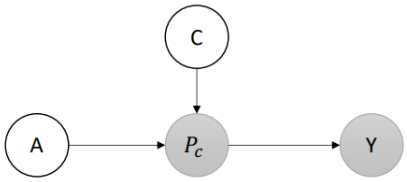
Graphical notation: elements

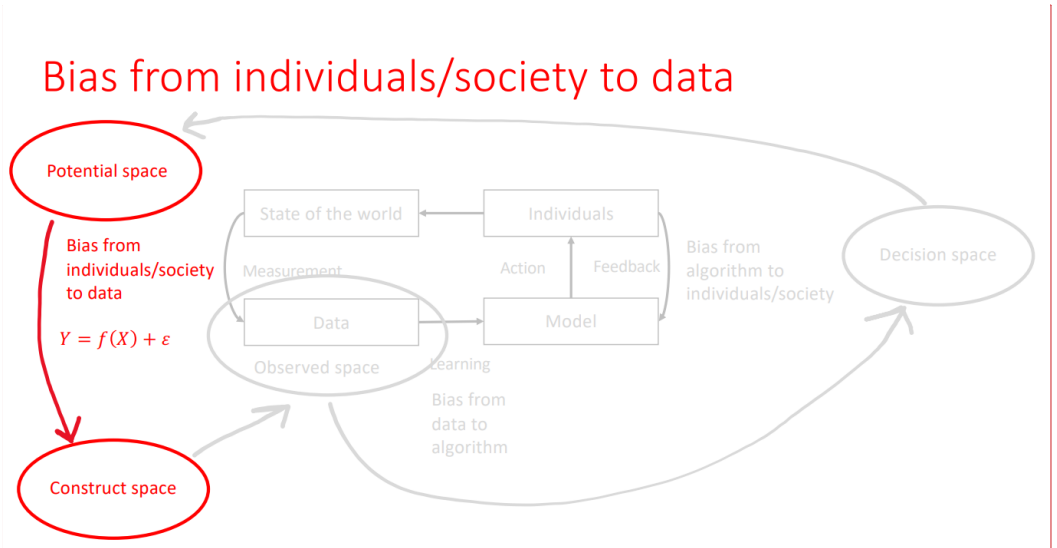
- Variables: circles
  - Grey circle: variable employed in the model f',
  - White circle: otherwise
- Dependence/correlation: a connecting arrow
- Example
  - a proxy for a capacity correlates with target variables,
  - and not-employed protected attribute correlates with the proxy

Commented [14]: متغیرها به شکل دایره هستند
gray: if it is used, is employed in model F.
White: whether it's not used, but it is correlated, connected to a certain other variables

Commented [15]: اگر A را به عنوان ویژگی جنسیت در نظر بگیریم و بر اساس مثال به دلایلی به پروکسی(اگر در نظر بگیریم ظرفیت یک درس باشد) متصل هست یا وابسته است (پروکسی را ضریب هوشی فرض کنیم)

Commented [16R15]: if there is a dependence, a correlation between let's say a proxy of a capacity and of course its capacity and a protected attribute, then we will write in this way. This means that A, let's say gender is not tracked, but we know that for some reason is connected, dependent or correlated to a proxy for example, intelligent quotient of have a capacity that again, we are not able to track and we are using a proxy of it. Both variables are dependent, are connected with P or C, which in turns is connected to the target variable. This is how to read this graph





Commented [17]: به این نوع بایاس که از potential space به construct space هست میگوییم Historical bias و میتواند شامل تمام نابرابری ها هستند(all the inequalities) روی پدیده های خاص تاثیر بگذارد.

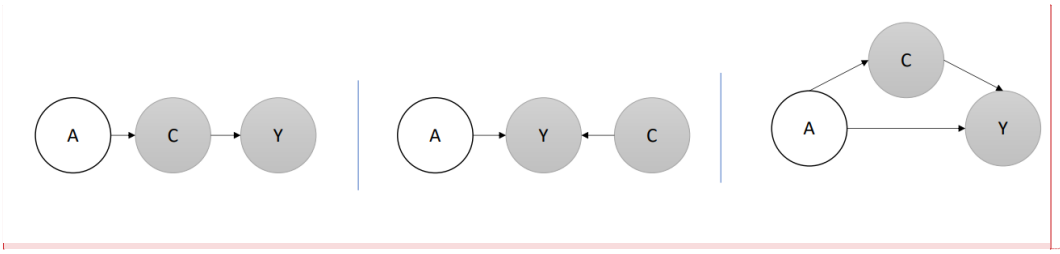
Historical bias

Inequalities and disproportions in the world:

- 95% of Fortune 500 CEO are men,
- different average income between men and women

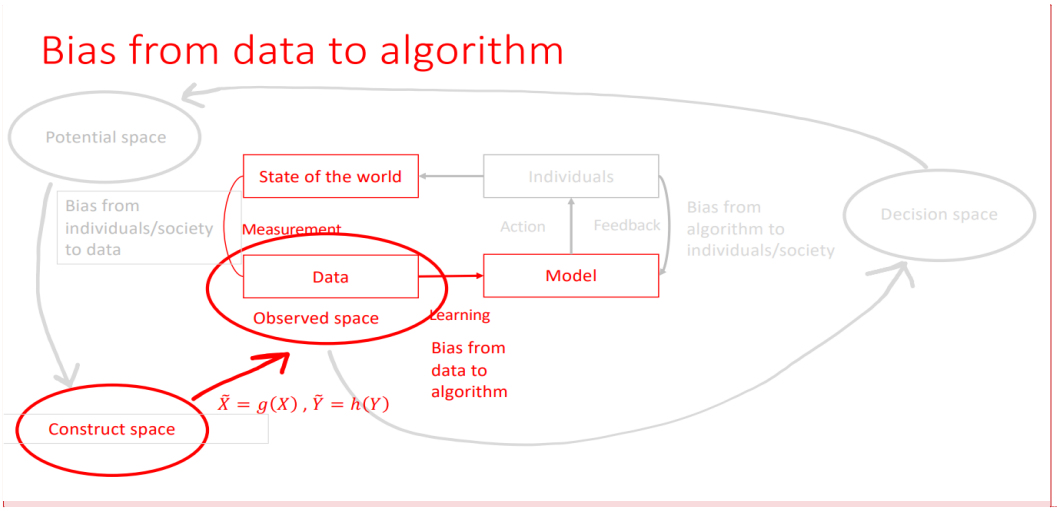
It occurs when:

- A relevant capacity variable is dependent on a protected attribute, and/or
- the target is dependent on a protect attribute



Commented [18]: به عنوان مثال، ما می دانیم که 95٪ از مدیرعاملان مرد هستند یا اکثریت مدیر عامل های اصلی مرد هستند. و این چیزی است که بر توانایی یک زن برای قرار گرفتن در پست های بالای یک شرکت تأثیر می گذارد.(جنسیت)
دقیقا زمانی اتفاق میافتد که شما میخواهید برای مثال احتمال اینکه یک فرد ظرف مدت 5 سال آینده مدیر شود را پیش بینی کنید با این واقعیت ظرفیت مربوط به یک ویژگی محافظت شده (Protected attribute)در این مثال جنسیت(وابسته می شود).
So you will have a bias and you will have that the target variable and all predictions are so dependent on the protected attribute

Commented [19]: But these are all situations in which historical bias can occur and can have an impact on your prediction classification algorithm.
این جند حالت را داریم historical برای بایاس
1. کپاسیتی به ویژگی وابسته باشه
2. تارگت ولیو به ویژگی وابسته باشه
3. هر دو به ویژگی وابسته باشند



Commented [20]: به این نوع بایاس که از construct space به observable space است میگوییم measurement که در زمان جمعآوری داده ها اتفاق می افتد و خیلی bias شبیه به قبلی است (historical bias) اما تفاوتش این است که در نابرابری ساختاری دخالت ندارد، بلکه بیشتر به فرآیند اندازه گیری مرتبط است.

Measurement bias

Similar to historical bias, but it does not involve the phenomenon itself Possible causes:

- Feature selection influenced by implicit mental models/subjective choices
- Granularity and quality of data varies across groups
- Although protected attributes are not used, their proxies are relevant

گرچه از ویژگی های محافظت شده استفاده نمی شود، پراکسی های آنها مرتبط هستند

Commented [21]: It occurs also when you have a quality of data that's might vary across specific groups. برای مثال اطلاعاتی که در زمینه ها مختلف بدست می آید بسته به گروهای خاص دقیق تر است

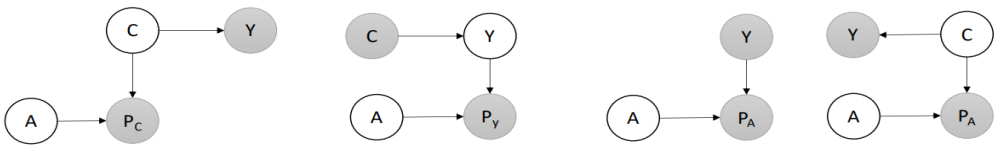
Measurement bias

It occurs when:

- a proxy of some capacity relevant to the target is employed, and that proxy is dependent on some sensitive characteristics;
- a proxy for the target is used, that is dependent on a protected attribute.
- a proxy of a protected attribute is used, that is related to the target and/or the capacity

Commented [22]: So you are using a proxy for a given capacity relevant to the target, and the proxy is connected to the protected attribute. Or you cannot access well your prediction, the construct of your prediction, you use a proxy, which is however, that is connected to a protected attribute.

Commented [23]: you have represented other possible situations where you have a proxy of a protected attribute that is directly connected to the target variable. And then discrimination can occur because you're using a proxy of a protected attribute. برای مثال اگر از کد پستی استفاده می کنید، این یک پروکسی رایج برای وضعیت اقتصادی است.



Commented [24]: گزارش جنایات خشن و خیابانی معمولاً دقیق تر از جرایم جزئی است و به شدت با فقر مرتبط است.(connected to socioeconomic conditions)
- داده های ترافیک برای مناطق شهری از جزئیات بیشتر نسبت به مناطق روستایی.
- داده های اجتماعی-اقتصادی که از نظر جغرافیایی دلالت دارند
- ضریب هوشی (ضریب هوش) با وضعیت اجتماعی-اقتصادی مرتبط است و به خوبی نشان دهنده هوش نیست.

Measurement bias: examples

- Reporting for violent, street crimes is usually more accurate than for minor crimes, and highly related to poverty

- Traffic data for urban areas richer of details than for rural areas

- Socio-economic data that are geographically connotated

- IQ (intelligence quotient) correlated to socio-economic status, and it does not well represent intelligence

Example 1: college admission

• A university uses an ML program to determine which candidates are more suitable for a degree program.

• SAT scores are a relevant variable used in the model

• Disparate classifications on protected attributes are observed.

• Two alternative assumptions:

  • a) SAT scores are a faithful representation of applicants' skills and competencies.

  • b) SAT scores do not faithfully represent applicants' skills and competencies

Example 1: college admission / scenario a

• a) SAT scores are a faithful representation of applicants' skills and competencies.

• Disparities are a consequence of some form of historical bias on C, impacting the actual skills and competencies of applicants

• $Y = f(C, Q) + \epsilon, C = C(A), Q \perp A$

Example 1: college admission / scenario b

Commented [25]: It means that the SAT scores are your capacity and you use it to make some prediction

Commented [26]: اگر اینجا discrimination وجود داشته باشد بخاطر وابستگی capacity به A هست.
مثال کشور افغانستان، می دانید که رفتن زنان به سطوح بسیار بالای آموزشی بسیار دشوار و حتی در دوره های خاصی از تاریخ دراین کشور غیرممکن است. بنابراین شما می دانید که یک historical bias وجود دارد، یک نابرابری potential وجود دارد، از (structural inequality)ساختاری که شرایط را برای جنس construct space گرفته تا space ماده قادر نمی‌سازد تا به بالاترین سطح آموزشی برود. بنابراین به عنوان یک نتیجه، مجموعه داده ای خواهید داشت که به شدت نامتعادل است. در هر صورت، تبعیض، تبیین تبعیض در historical bias است.

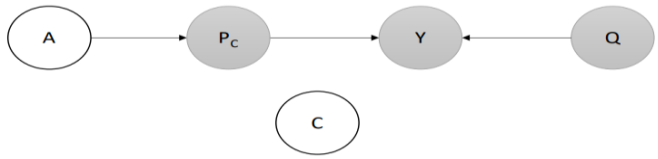
Commented [27R26]: If you have a discrimination, then the discrimination of course, because of the capacity, so the SAT score is directly connected to the protected attribute. So if there is a discrimination, it occurs mostly because of data.

Commented [28]: the target variable is a function of the competencies, the skills given measured by the SAT score(C) and some other variable.(Q)

• b) SAT scores do not faithfully represent applicants' skills and competencies

• SAT score disparities are the result of a measurement bias on C: SAT score is not the proper way to assess skills and competencies

• $Y = f(P_c, Q) + \epsilon, C, Q \perp A$



Example 2: loan applications

• A bank uses an ML model to determine creditworthiness, i.e. whether loan applications should be approved or denied.

• It is possible to track repayment rates, and it is observed that the rate of repayment heavily varies with respect to gender.

• Two scenarios,

• a) historical bias on Y , i.e. the repayment rate disparity reflects a real mismatch in creditworthiness between men and women;

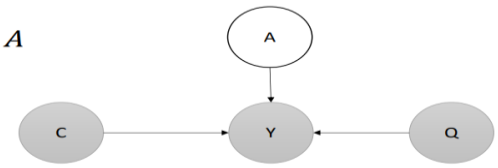
• b) measurement bias on Y , i.e. the observed repayments are a skewed measure of real creditworthiness.

Example 2: loan applications / scenario a

• a) historical bias on Y , i.e., the repayment rate disparity reflects a real mismatch in creditworthiness between men and women;

• It is a consequence of a structural discrimination, e.g., income disparities between men and women

• $Y = f(C, Q, A) + \epsilon, C, Q \perp A$



Example 2: loan applications / scenario b

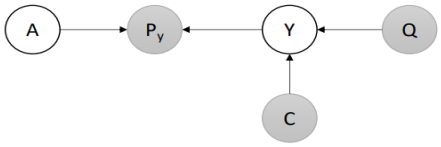
Commented [29]: In this case, when you have a discrimination, you have it, not because of historical bias, but because of a measurement bias, because it occurs on the proxy. The relation is with the proxy, not with the competence (C)
مثالی که اینجا هست این است که اگر بخواهیم میزان ساعات کاری یا تجربه کاری افراد را اندازه گیری کنیم
میدانیم که خانومها به طور میانگین کمتر از مردها کار میکنن بخاطر اینکه باید به مرخصی هایی مثل زایمان یا مراقبت از فرزند یا سالمندان بروند.
معرفی proxy را به عنوان experiance اگر درنظر بگیرید که کنیم، اما این پروکسی، میزان زمان کار، تعداد ساعات کار، ممکن است به دلایلی که گفته شد بسیار با جنسیت مرتبط باشد، زیرا به خصوص در یک محدوده سنی معین، زنان تعهدات مراقبت بیشتری دارند.

Commented [30]: اگر فرض کنید که نرخ بازپرداخت معیاری برای سنجش ارزش اعتبار است، آنگاه وقتی تبعیض رخ می دهد، به دلیل شکاف جنسیتی است. چون به عنوان یک زن درآمد شما کمتر است، و بنابراین ممکن است مشکلات بیشتری در بازپرداخت بدهی خود پیدا کنید.
این نتیجه یک تبعیض ساختاری است،بخاطر، تفاوت درآمد بین مردان و زنان

Commented [31]: you are not using the real credit worthiness. You are using a proxy of it. And the proxy is correlated. So capability to repay is a proxy of credit worthiness. and is correlated the proxy with the protected attributes. The results is the same. You will get a discrimination.

• b) measurement bias on Y , i.e., the observed repayments are a skewed measure of real creditworthiness

- $Y = f(C, Q) + \epsilon, C, Q \perp A$, and the observed dependence on the protected attribute comes from the proxy P_Y



Representation bias

Representation bias occurs when data are not representative of the actual population.

- Sampling method does not reach all N equally

- Changes in the population not detected

Representation bias is sufficient to create discrimination

Representation bias - examples

- Training with images taken from specific geographic locations

- Behavior of people that changes over time: e.g., Covid-19 lockdowns disrupted validity of several prediction models

Omitted relevant variables bias

• It may occur when a variable relevant to the target/goal is omitted and not present in the collected data.

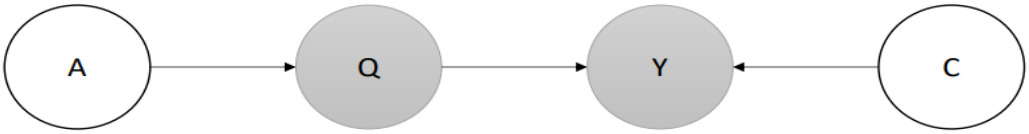
• If the other variables present in the dataset have some dependence on protected attributes, the trained model will learn those dependencies and outcomes will be affected by spurious dependence on sensitive attributes.

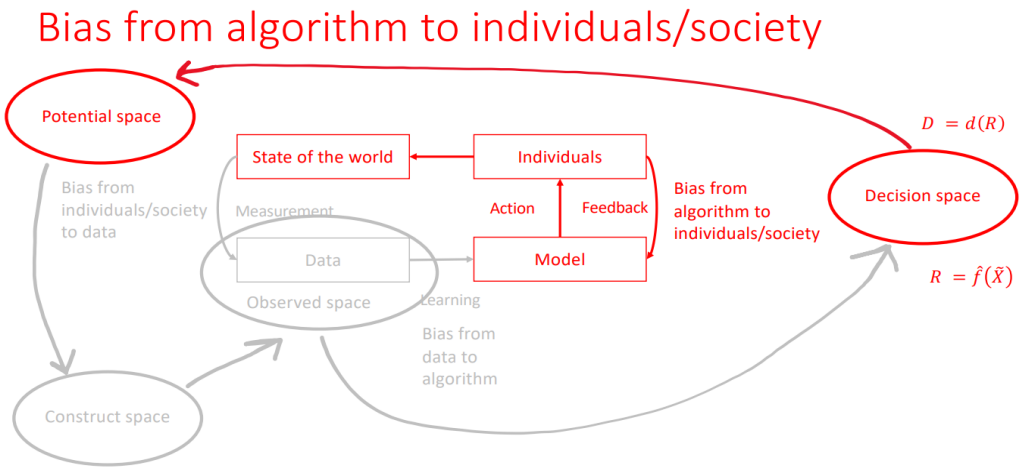
Commented [32]: زمانی اتفاق می افتد که نمونه های گرفته شده همه افراد جامعته یا شرایط مختلف را در نظر نگرفته باشد یا داده ای از آن بخش خاص در دسترس نباشد.
مثلا استفاده از گوشی هوشمند به طور متوسط بین جوانان و افراد مسن بسیار متفاوت است.
The presentation bias explains alone is sufficient to create a discrimination.
بنابراین ممکن است target variable (Y) مستقیماً به یک protected attribute(A) یا از طریق یک ظرفیت خاص (C) یا به نوبه خود به یک متغیر دیگر (Q) به نوبه خود به protected attribute متصل باشد.

Commented [33]: آموزش با تصاویر گرفته شده از مکان های جغرافیایی خاص.
کاملاً بدیهی است که اگر من فقط از تصاویر افرادی از شرق آسیا استفاده کنم، الگوریتم با چهره های آفریقایی به خوبی کار نمی کند.

Commented [34]: من الگوریتم‌های زیادی را مشاهده می‌کردم که رفتار مردم در طول همه‌گیری کرونا کاملاً تغییر کرد. recommandations، ، advertising recommendation، الگوریتم‌ها در طول همه‌گیری کار نمی‌کردند زیرا عادات ما کاملاً مختل شده بود..
خاص capacity ممکن است موقعیتی داشته باشید که مثلاً یک اصلاً ردیابی نشود

• Omission of a relevant variable alone cannot be a source of disparities, but it can amplify existing biases (e.g. historical biases)



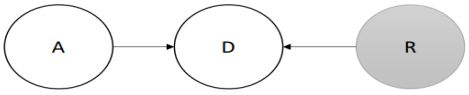


Deployment bias

• It arises when the decision-making process -based on the algorithm's predictions/classifications- has harmful downstream consequences (e.g., as seen in predictive policing).

• It can be seen as a bias going from algorithm to the society through decision makers.

• Examples:

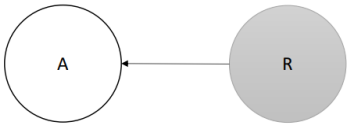


Algorithmic bias

• When algorithmic outcomes affect the behavior of people (although data was not biased), exacerbating performance disparities on underrepresented groups

- Aggregation bias

- Learning bias
- Evaluation bias



Aggregation bias and learning bias

- **Aggregation bias:**

- it arises when subgroups are so different that different ML models should be used instead of only one for everyone;

- it results in inconsistency in mapping inputs to labels, i.e., a different probability of receiving a label given some features.

- **Learning bias**

- It arises when algorithmic design choices (e.g., the learning objective function) are not equally suited for all subgroups.

Commented [35]: زمانی به وجود می آید که زیرگروه ها آنقدر متفاوت هستند که باید به جای یک مدل برای همه از مدل های مختلف ML استفاده شود.

Commented [36]: منجر به ناسازگاری در نگاشت ورودی ها به برچسب ها می شود، به عنوان مثال، با توجه به برخی ویژگی ها، احتمال متفاوت دریافت برچسب وجود دارد.

Commented [37]: زمانی به وجود می آید که انتخاب های طراحی الگوریتمی (مثلاً تابع هدف یادگیری) برای همه زیرگروه ها به یک اندازه مناسب نباشد.

Aggregation bias and learning bias: examples

- Clinical aid tools might need different models for different groups of people

- Slangs in social network used by specific groups of people, but only one language model used

- On web platforms, when later review rates are strongly influenced by previous ones, they should be treated separately

Commented [38]: مثال الگوریتم‌های پیش‌بینی سرطان پوست یا هر ابزار بالینی دیگری که ممکن است برای زیر گروه‌های مختلف جمعیت مناسب باشد.مدل باید متفاوت باشد زیرا پوسته ها بسیار متفاوت است.
you cannot use the same model for skin cancer prediction on a population of white skin and on a population of black skin.

Commented [39]: مثال در مورد review rate که وقتی یک رستوران یا محصولی را بررسی می‌کنید، اگر نظرات قبلی را ببینید، ممکن است به نحوی تحت تأثیر قرار بگیرید.

Evaluation bias It arises when:

- evaluation data do not well represent the target population, or training and evaluation/operation data are very different;

- model performance metrics used are poorly relevant for the relevant target population and the application context

Commented [40]: وقتی مدل خود را آموزش می‌دهید، آن را ارزیابی می‌کنید، آن را آزمایش می‌کنید و سپس از زمینه‌ای کاملاً متفاوت استفاده می‌کنید. مثلا مدل با سابقه خرید در سوپرمارکت‌های آمریکایی آموزش دیده و سپس با داده‌های سوپرمارکت‌های آفریقایی و نمونه‌هایی ارزیابی شده است

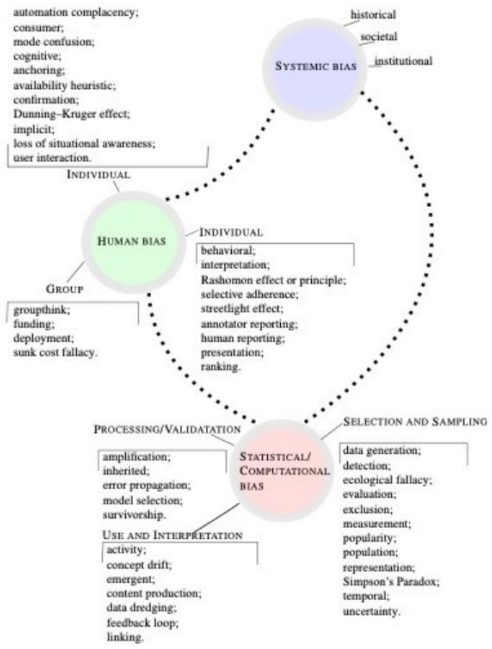
Evaluation bias: examples

- facial recognition mostly trained on white face, then used/evaluated in contexts where other skin colors exist;

- a model trained with purchases record on American supermarkets then evaluated with data from African supermarkets.

- Model minimizes false negative, when false positive minimization is more important

A more general view on bias



Source:Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence*. NIST Special publication 1270

Chapter 05-Formalization of Algorithmic Fairness

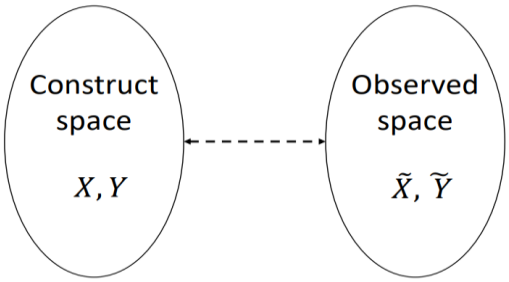
Fair means... be to have an impact that is the same on different people, on different groups of people.

One of the very first desired characteristics of an algorithm in order to be fair is that it classifies given social groups in the same way.

Formalization

- Y : target variable, the quantity to be predicted
- X: features variables that determine Y
- Phenomenon of interest modeled by the relationship Y= f(x)+e

Space of observations



Space of observations

Commented [41]: درامتحان به این صورت است
Possible cases of the exam might require you to compute this, one of these criteria and to explain them. So to check whether there is a discrimination occurring and explain it, for example. or to choose one algorithm or the other one with respect to a given criteria of fairness, of your choice. So you choose a criteria, you explain why you choose it, and then you say, okay, algorithm one is preferable instead of algorithm two.

Commented [42]: برای مثال احتمال پیش بینی یه برنامه درسی برای سفیدپوستان و سیاه پوستان یا مردان و زنان باید یکسان باشد
the probability of predicting a good curriculum should be the same for whites, for black

- Construct space: $Y = f(X) + \varepsilon$
- Data collection→ Space of observations:
  - $\tilde{X} = g(X)$
  - $\tilde{Y} = h(Y)$
- Data from observations: $(\tilde{x}_1, \tilde{y}_1)$, $(\tilde{x}_2, \tilde{y}_2)$, $(\tilde{x}_n, \tilde{y}_n)$
  - Each pair is
    - An instance $\tilde{x}_i$(a vector)
    - A label $\tilde{y}_i$ (in most cases, we will assume $\tilde{y}_i \in \{0,1\}$
- A feature that is a protected attribute is named A

Protected attributes

We will consider those from Article 21 - Non-discrimination EU Charter of Fundamental Rights

1. Any discrimination based on any ground such as **sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation** shall be prohibited.

2. Within the scope of application of the Treaties and without prejudice to any of their specific provisions, any discrimination on grounds of nationality shall be prohibited.

Predictions and decisions

- $\widehat{f} \sim f$ estimate of f
- $R = \hat{f}(\tilde{X})$ prediction R made by the classifier trained on $\tilde{X}$
- $D = d(R)$ decision D made with a decision rule d applied on R
- In certain applications, it might be decided that $D = R$
- or that a decision relies upon:
  - additional environmental information Z
  - considerations about a protected attribute A (or more)
- → $D = d(R, A, Z)$

Classification criteria

• Accuracy of a classifier is defined as P(Y=R)

• When classification is binary, it is possible to use the notation of conditional probability

P(classification | Condition) Where:

P(classification | Condition) = P(classification **conjuction** Condition) /P(condition)

Classification criteria

Classification	Condition	Notion
$R = 1$	$Y = 1$	True positive (Right inference)
$R = 0$	$Y = 1$	False negative (Error type II)
$R = 1$	$Y = 0$	False positive (Error type I)
$R = 0$	$Y = 0$	True negative (Right inference)

Fairness criteria

Criteria

• Independence

given protected attributes, we may require that an algorithm has the same probability of giving either an opportunity in the case of assigning an opportunity in the case of the job application or giving a penalty, a risk score is a potential penalty. And this first criterion is called independence.

• Separation

• Sufficiency:

for a given level of the classification, so if the classification is positive, we expect a higher presence of actually positive instances than in the group of negative classifications.

Commented [43]: در این مرحله چیزی که میخواهیم این است که
our positive classification to be the same
بنابراین اگر الگوریتم classification errors نباید دارد، خطاهای بیشتری برای مردان و خطاهای کمتری برای زنان ایجاد کند.
الگوریتم های تشخیص چهره عملکرد بسیار بدی با چهره های زنان سیاه پوست دارند.
this is an example. of different error rates for different social groups. Black women is a particular case of social group.

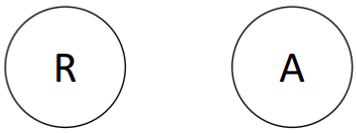
Commented [44]: در این مرحله کیفیت طبقه بندی ها را بررسی می شود

Commented [45]: در این مرحله میخواهیم که الگوریتم ما به خوبی کالیبره شده باشد

Independence

$$R \perp A$$

$$\mathbb{P}\{R = 1|A = a\} = \mathbb{P}\{R = 1|A = b\}$$



Independence - relaxed

$$R \perp A$$

$$\frac{\mathbb{P}\{R = 1|A = a\}}{\mathbb{P}\{R = 1|A = b\}} \geq 1 - \epsilon$$

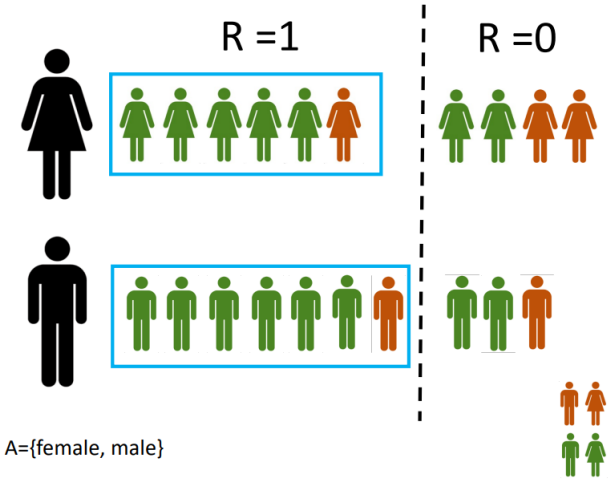
Independence – example 1

- **Independence NOT respected**

Commented [46]: Independence requires that the classification is independent of the protected attributes.

Commented [47]: برای امتحان:
So usually at the exam, I ask only for a given fairness criteria, or I ask you to select a fairness criteria. It is very, very important that you don't show me only the result, that you show me at least for a computation the steps. Why? Because If you make only a computation error, I don't count it. But if you make a conceptual error, I will count it. And I cannot know whether you made just a computation error or a conceptual error if I see only the final result.

Independence – example 1



Gender	R=1	R=0	tot
Female	6	4	10
Male	7	3	10
tot	13	7	20

$$\mathbb{P}\{R=1|A=f\}=0.60$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{A=f\}}{\mathbb{P}\{A=f\}}=\frac{\frac{6}{20}}{\frac{10}{20}}=0.60$$

$$\mathbb{P}\{R=1|A=m\}=0.70$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{A=m\}}{\mathbb{P}\{A=m\}}=\frac{\frac{7}{20}}{\frac{10}{20}}=0.70$$

Wrongly classified

Correctly classified

Commented [48]: The green symbols are the correctly classified elements, people.
The wrong classifications are red.
برای نمایش استفاده میکنیم confusion matrix از

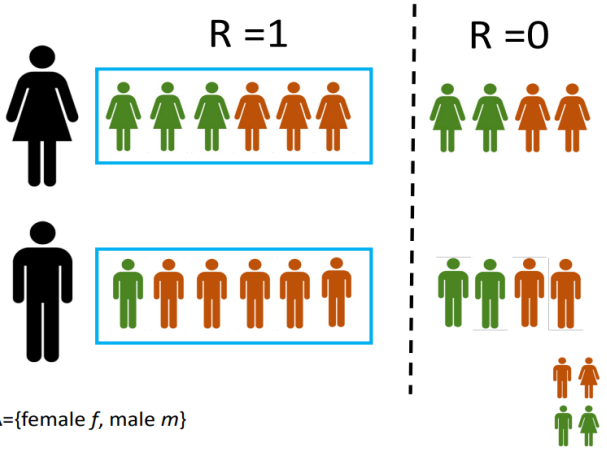
remember that independence, does not take into account the quality of the classification. So we don't track for independence the errors. And this is, of course, this is a limitation of independence. It's the very first criteria that you can put in place for requiring a fair algorithm classified with a positive instance in the same way all the social groups.

در این مثال استقلال در نظر گرفته نشده بود میتوان دید که اختلاف درصد زنان و مردان 10 درصد شده است مگر اینکه در صورت سوال گفته شده باشد یه ترشولد خطلا مثلا 10 درصدی داریم(**consider relaxation of up to** 10%).

Independence – example 2

- **Independence respected**

Independence – example 2



Gender	R=1	R=0	tot
Female	6	4	10
Male	6	4	10
tot	13	7	20

$$\mathbb{P}\{R=1|A=f\}=0.60$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{A=f\}}{\mathbb{P}\{A=f\}}=\frac{\frac{6}{20}}{\frac{10}{20}}=0.60$$

$$\mathbb{P}\{R=1|A=m\}=0.60$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{A=m\}}{\mathbb{P}\{A=m\}}=\frac{\frac{6}{20}}{\frac{10}{20}}=0.60$$

So I might ask you, take independence and select which algorithm is preferable. And you will choose, of course, the one that is with lowest differences in terms of independence. And then I might ask you which limitation you still observe in the algorithm that you selected as most preferable. One of the possible limitations could be this one, that even though if algorithm, let's say one, has the lowest difference between probability of R equal one for female and males, but still expose very higher rates(نسبت خطاها در زنان که حددو 50 درصد هست و مردان 90 درصد وقتی r=1 است). This is a common limitation that you might find even when you select an algorithm as preferable.

Commented [49]: در امتحان این موارد ممکن است پرسیده شود

Achieving Independence

• Pre-processing: Adjust the feature space to be uncorrelated with the sensitive attribute.

• At training time: Work the constraint into the optimization process that constructs a classifier from training data.

• Post-processing: Adjust a learned classifier so as to be uncorrelated with the sensitive attribute.

Independence: pros and cons

• Advantages:

  • it can be applied at every stage of the process

•Disadvantages:

- It ignores the possible correlation between Y and A.
- It allows to have good classifications in one group, and random classifications in another

Criteria

- Independence
- Separation
- Sufficiency

Commented [50]: همونطور که گفتیم استقلال یه ضعف رو در classification پیش بینی ها و quality داشت اونم اینکه نپر نمیگرفت پس اگر برامون مهم باشه باید از معیار دیگری مثل Separation استفاده کنیم

Separation

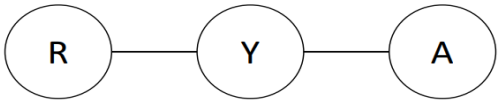
Commented [51]: Separation requires that given a certain level of the target attributes, the classification is independent on the protected attributes

The separation criteria has two requirements(دو خط پایین)

$$R \perp A \mid Y$$

$$\mathbb{P}\{R = 1 | Y = 1, A = a\} = \mathbb{P}\{R = 1 | Y = 1, A = b\}$$

$$\mathbb{P}\{R = 1 | Y = 0, A = a\} = \mathbb{P}\{R = 1 | Y = 0, A = b\}$$



The separation criterion is a criterion of equality of error rates

Classification criteria

Event	Condition	Notion
$\hat{Y} = 1$	$Y = 1$	True positive (Right inference)
$\hat{Y} = 0$	$Y = 1$	False negative (Error type II)
$\hat{Y} = 1$	$Y = 0$	False positive (Error type I)
$\hat{Y} = 0$	$Y = 0$	True negative (Right inference)

Separation

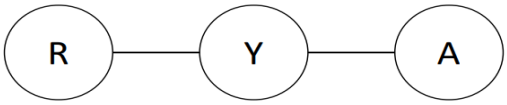
True positive

False positive

$$R \perp A \mid Y$$

$$\mathbb{P}\{R = 1 \mid Y = 1, A = a\} = \mathbb{P}\{R = 1 \mid Y = 1, A = b\}$$

$$\mathbb{P}\{R = 1 \mid Y = 0, A = a\} = \mathbb{P}\{R = 1 \mid Y = 0, A = b\}$$



The separation criterion is a criterion of equality of error rates

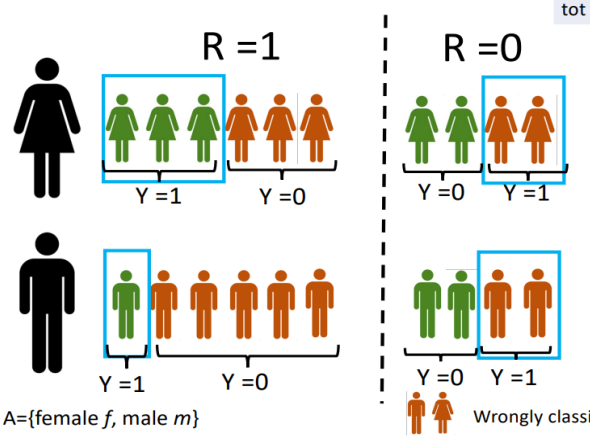
Separation – example 1

- **Separation NOT respected**

Commented [52]: محاسبه میشوند R=1 هر دو بر روی برای امتحان
Please do not place into your computation false negatives when computing separation. This is an error. Separation requires only true positive, equalization of true positive and false positives.
فقط اگر گفته بود چه محدودیت بیشتری را میبینید باید محاسبه شود
Which further limitation you still observe? Then you could compute the false negatives and say, OK, there is a high rate of false negatives. It's a further limitation, but not as a... requirement of separation.

Separation – example 1 / Y=1

Gender	Y=1			Y=0			TOT
	R=1	R=0	*tot*	R=1	R=0	*tot*	
Female	3	2	*5*	3	2	*5*	10
Male	1	2	*3*	5	2	*7*	10
tot	4	4	*8*	8	4	*12*	20



$$\mathbb{P}\{R=1\,|Y=1, A=f\} = 0.60$$

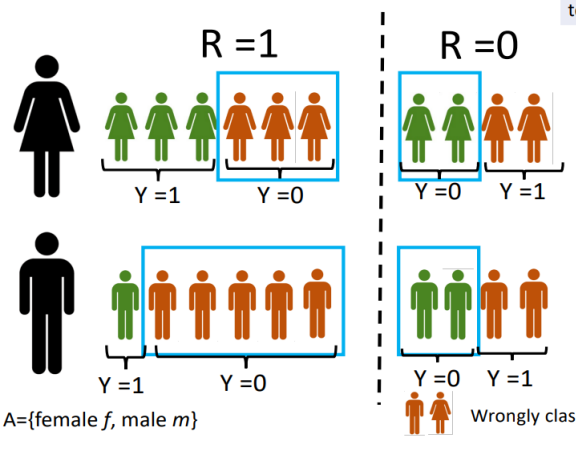
$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{Y=1,A=f\}}{\mathbb{P}\{Y=1,A=f\}} = \frac{\frac{3}{20}}{\frac{5}{20}} = 0.60$$

$$\mathbb{P}\{R=1|Y=1, A=m\} = 0.33$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{Y=1,A=m\}}{\mathbb{P}\{Y=1,A=m\}} = \frac{\frac{1}{20}}{\frac{3}{20}} = 0.33$$

Separation – example 1 / Y=0

Gender	Y=1			Y=0			TOT
	R=1	R=0	*tot*	R=1	R=0	*tot*	
Female	3	2	*5*	3	2	*5*	10
Male	1	2	*3*	5	2	*7*	10
tot	4	4	*8*	8	4	*12*	20



$$\mathbb{P}\{R=1\,|Y=0, A=f\} = 0.60$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{Y=0,A=f\}}{\mathbb{P}\{Y=0,A=f\}} = \frac{\frac{3}{20}}{\frac{5}{20}} = 0.60$$

$$\mathbb{P}\{R=1|Y=0, A=m\} = 0.71$$

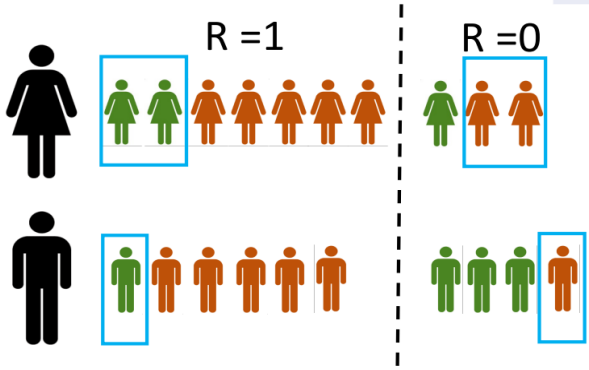
$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{Y=0,A=m\}}{\mathbb{P}\{Y=0,A=m\}} = \frac{\frac{5}{20}}{\frac{7}{20}} = 0.71$$

Separation – example 2

- **Separation PARTLY respected**

Separation – example 2 / Y=1

Gender	Y=1			Y=0			TOT
	R=1	R=0	*tot*	R=1	R=0	*tot*	
Female	2	2	*4*	5	1	*6*	10
Male	1	1	*2*	5	3	*8*	10
tot	3	3	*6*	10	4	*14*	20



$$\mathbb{P}\{R = 1 \mid Y = 1, A = f\} = 0.50$$

$$\frac{\mathbb{P}\{R=1\} \cap \mathbb{P}\{Y=1, A=f\}}{\mathbb{P}\{Y=1, A=f\}} = \frac{\frac{2}{20}}{\frac{4}{20}} = 0.50$$

$$\mathbb{P}\{R = 1 \mid Y = 1, A = m\} = 0.50$$

$$\frac{\mathbb{P}\{R=1\} \cap \mathbb{P}\{Y=1, A=m\}}{\mathbb{P}\{Y=1, A=m\}} = \frac{\frac{1}{20}}{\frac{2}{20}} = 0.50$$

A={female *f*, male *m*}

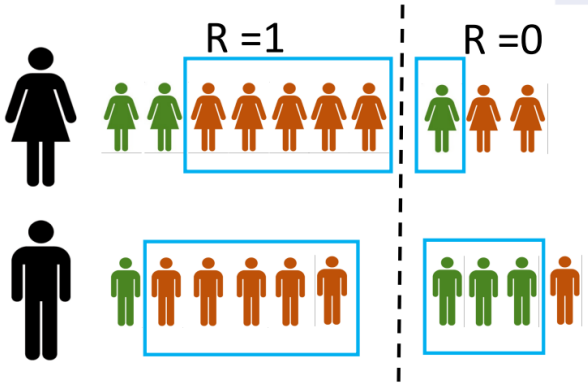
Wrongly classified

Correctly classified

Commented [53]: whenever only one criteria is respected, we say the separation is partially respected.

Separation – example 1 / Y=0

Gender	Y=1			Y=0			TOT
	R=1	R=0	*tot*	R=1	R=0	*tot*	
Female	2	2	*4*	5	1	*6*	10
Male	1	1	*2*	5	3	*8*	10
tot	3	3	*6*	10	4	*14*	20



$$\mathbb{P}\{R = 1 \mid Y = 0, A = f\} = 0.83$$

$$\frac{\mathbb{P}\{R=1\} \cap \mathbb{P}\{Y=0, A=f\}}{\mathbb{P}\{Y=0, A=f\}} = \frac{\frac{5}{20}}{\frac{6}{20}} = 0.83$$

$$\mathbb{P}\{R = 1 \mid Y = 0, A = m\} = 0.63$$

$$\frac{\mathbb{P}\{R=1\} \cap \mathbb{P}\{Y=0, A=m\}}{\mathbb{P}\{Y=0, A=m\}} = \frac{\frac{5}{20}}{\frac{8}{20}} = 0.63$$

A={female *f*, male *m*}

Wrongly classified

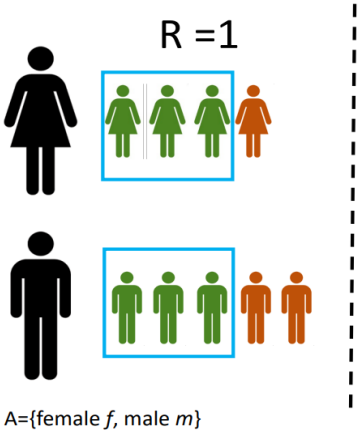
Correctly classified

Separation – example 3

- **Separation respected**

Separation – example 1 / Y=1

Gender	Y=1			Y=0			TOT
	R=1	R=0	*tot*	R=1	R=0	*tot*	
Female	3	2	*5*	1	1	*2*	7
Male	3	2	*5*	2	2	*4*	9
tot	6	4	*10*	3	3	*6*	16



$$\mathbb{P}\{R = 1 \,|Y = 1, A = f\} = 0.60$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{Y=1,A=f\}}{\mathbb{P}\{Y=1,A=f\}} = \frac{\frac{3}{20}}{\frac{5}{20}} = 0.60$$

$$\mathbb{P}\{R = 1|Y = 1, A = m\} = 0.60$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{Y=1,A=m\}}{\mathbb{P}\{Y=1,A=m\}} = \frac{\frac{3}{20}}{\frac{5}{20}} = 0.60$$

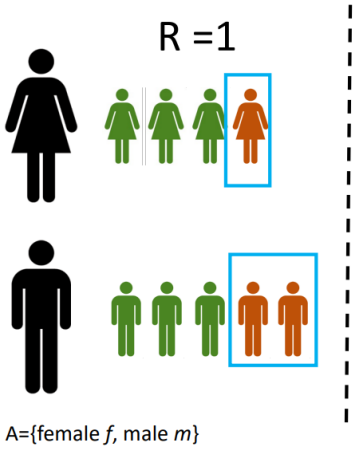
A={female *f*, male *m*}


Wrongly classified


Correctly classified

Separation – example 1 / Y=0

Gender	Y=1			Y=0			TOT
	R=1	R=0	*tot*	R=1	R=0	*tot*	
Female	3	2	*5*	1	1	*2*	7
Male	3	2	*5*	2	2	*4*	9
tot	6	4	*10*	3	3	*6*	16



$$\mathbb{P}\{R = 1 \,|Y = 0, A = f\} = 0.50$$

$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{Y=0,A=f\}}{\mathbb{P}\{Y=0,A=f\}} = \frac{\frac{1}{20}}{\frac{2}{20}} = 0.50$$

$$\mathbb{P}\{R = 1|Y = 0, A = m\} = 0.50$$

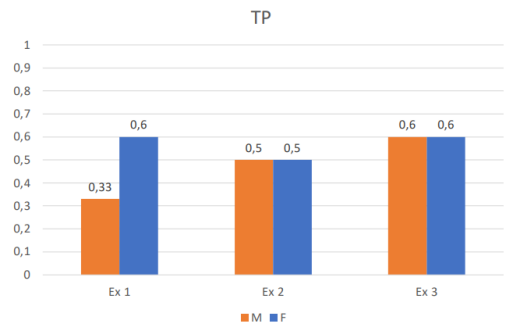
$$\frac{\mathbb{P}\{R=1\}\cap\mathbb{P}\{Y=0,A=m\}}{\mathbb{P}\{Y=0,A=m\}} = \frac{\frac{2}{20}}{\frac{4}{20}} = 0.50$$

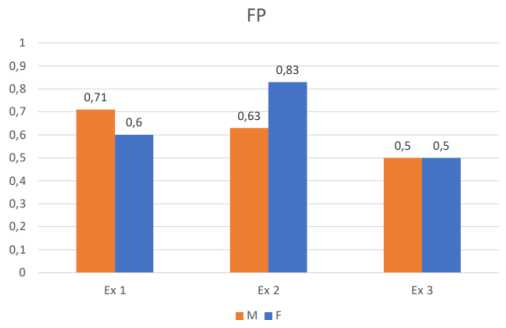
A={female *f*, male *m*}


Wrongly classified


Correctly classified

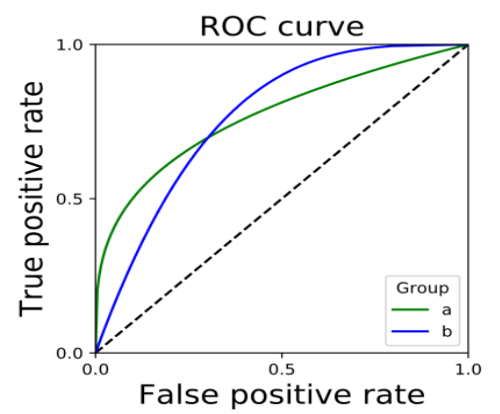
Separation - Summary of the 3 examples

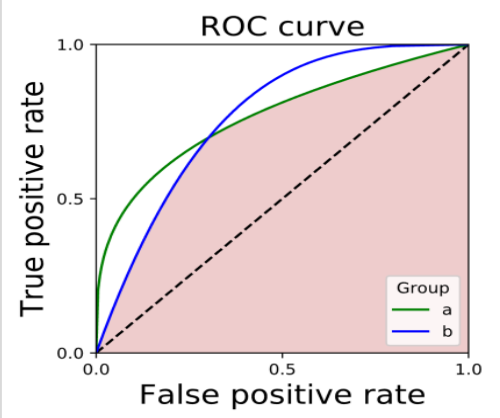




How to obtain separation

- Training : via ad-hoc optimization
- Post-elaboration: it is verified whether the intersection of the ROC curves of each group occurs





Separation: pros and cons

- **Advantages:**
  - It is compatible with R=Y
  - It incentives to reduce errors uniformly in all groups
- **Disadvantages:**
  - more difficult to apply
  - Does not take into account false negative rate

separation is much more strict. Requires also that you look at the quality of the predictions. **Not all errors are taken into account**. The **false negative rate is not taken into account**. And in

certain applications, the false negative rate might be more important the false positive rate. It depends on the **context of application**. So receiving an error as a false positive rate might be an influence for certain classifications. As I said, it's much more difficult to achieve it.

Criteria

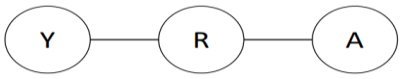
- Independence
- Separation
- Sufficiency

Commented [54]: Sufficiency is the reverse of the separation

Sufficiency

$$Y \perp A \mid R$$

$$\mathbb{P}\{Y = 1 | R = r, A = a\} = \mathbb{P}\{Y = 1 | R = r, A = b\}$$



The R score is sufficient to predict Y, the sensitive attribute is not needed in the model.

Classification criteria

Event	Condition	Notion
$Y = 1$	$R = 1$	Positive predictive value (PPV)
$Y = 1$	$R = 0$	False omission rate (FOR)

Sufficiency – binary case

Sufficiency – binary case

Positive predictive value

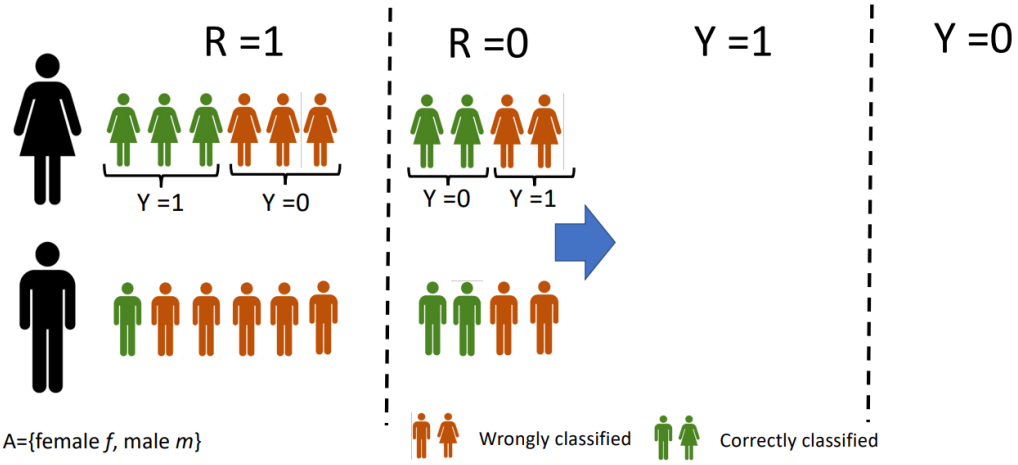
False omission rate

$$Y \perp A \mid R$$

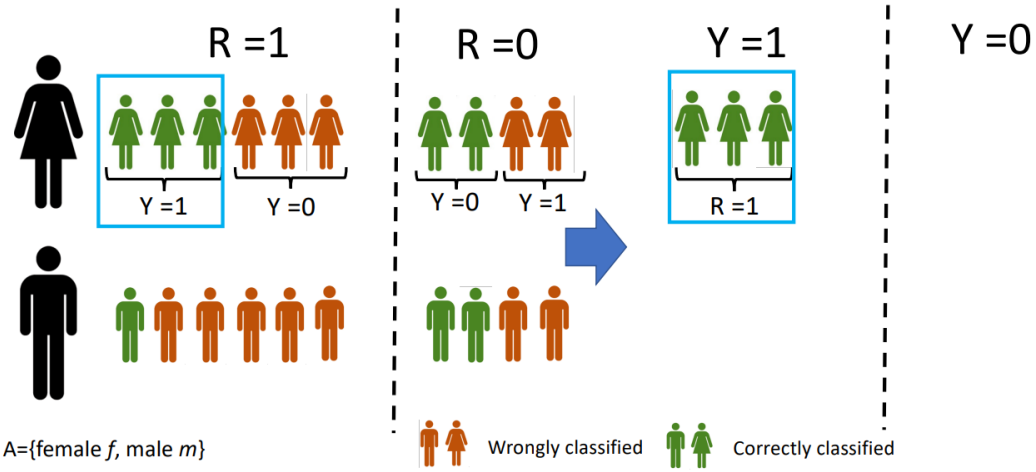
$$\mathbb{P}\{Y = 1 \mid R = 1, A = a\} = \mathbb{P}\{Y = 1 \mid R = 1, A = b\}$$

$$\mathbb{P}\{Y = 1 \mid R = 0, A = a\} = \mathbb{P}\{Y = 1 \mid R = 0, A = b\}$$

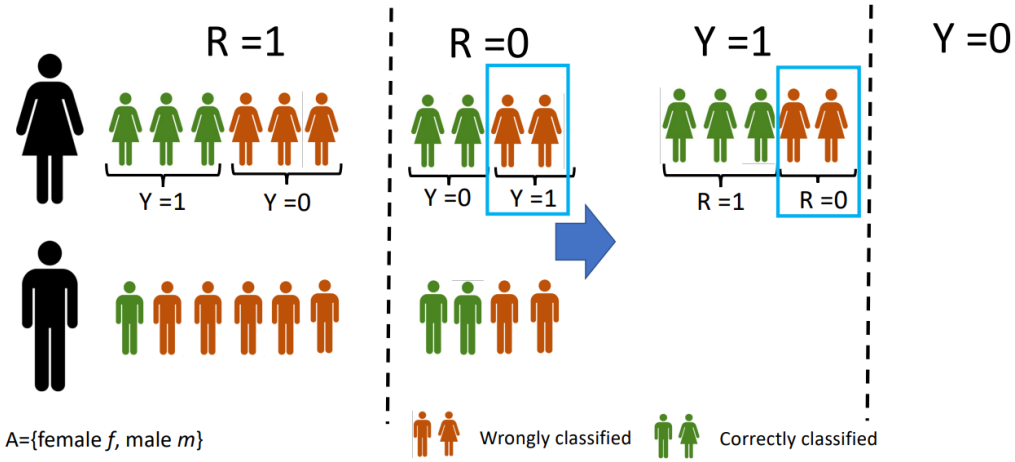
Separation – example 1 → Sufficiency example 1



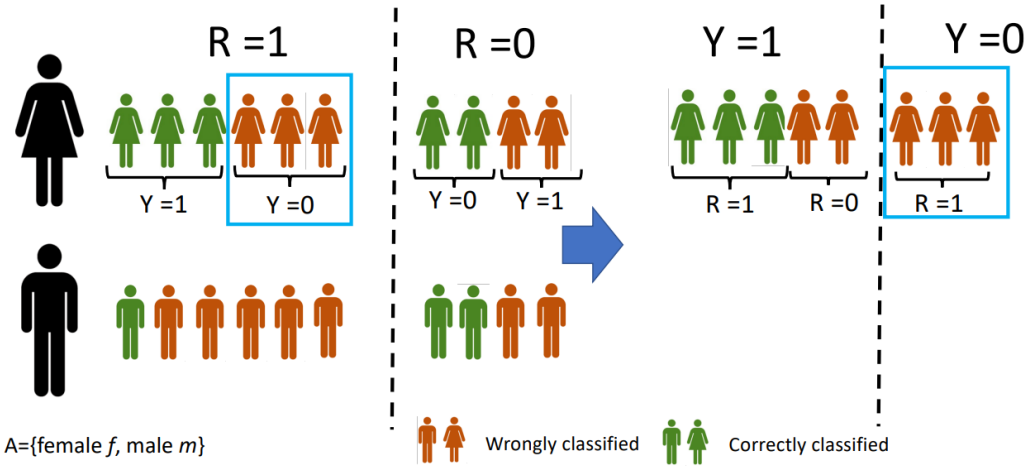
Separation – example 1 → Sufficiency example 1



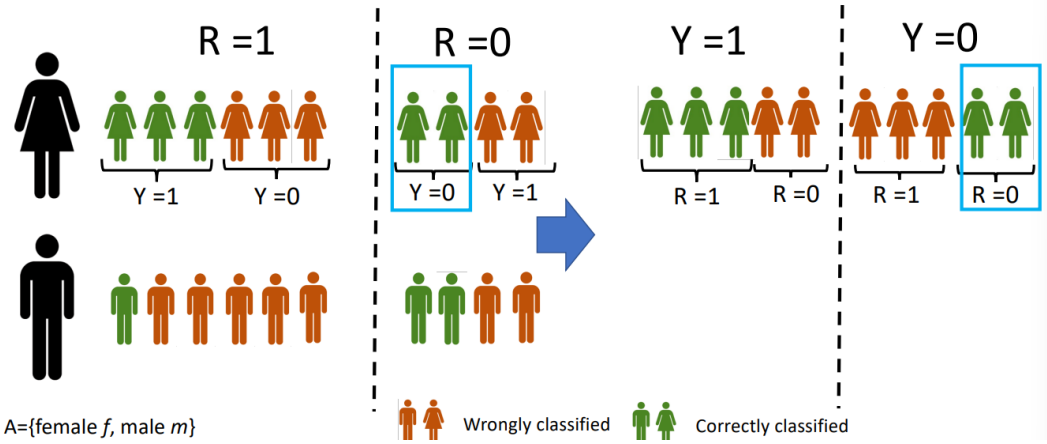
Separation – example 1 → Sufficiency example 1



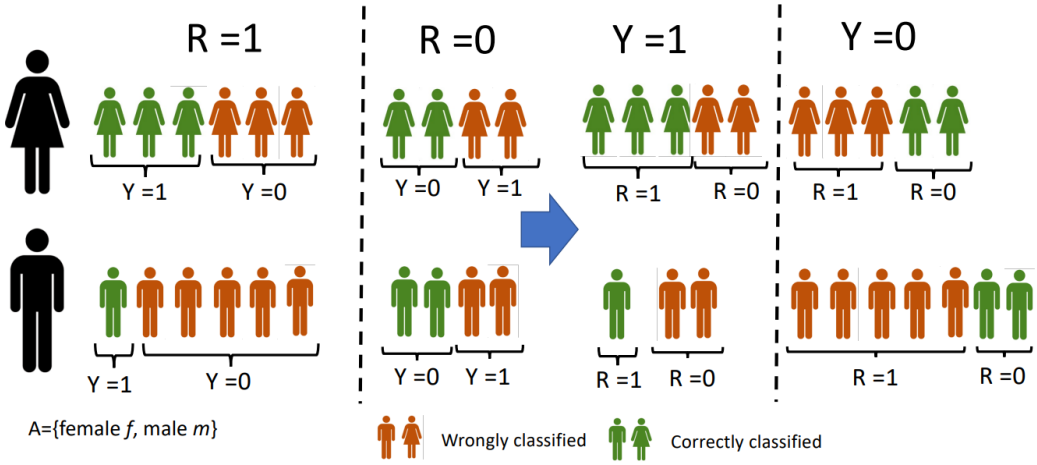
Separation – example 1 → Sufficiency example 1



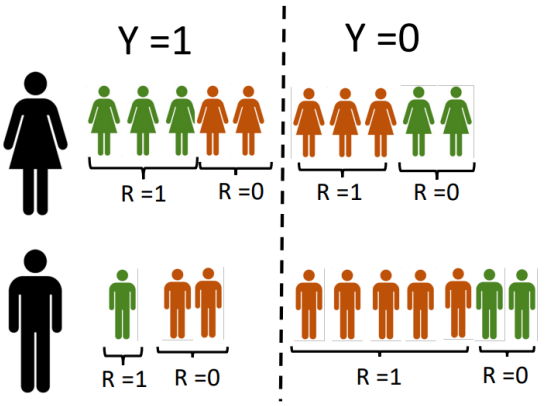
Separation – example 1 → Sufficiency example 1



Separation – example 1 → Sufficiency example 1



Separation – example 1 → Sufficiency example 1



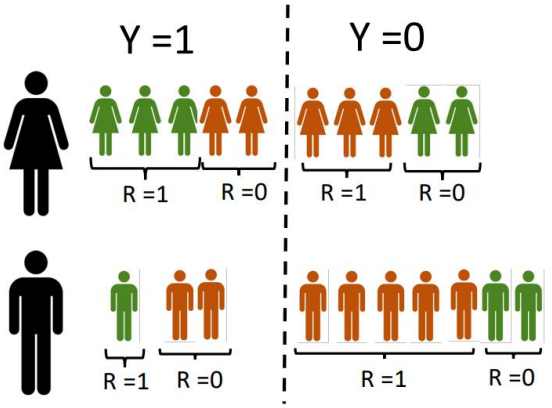
Gender	Y=1			Y=0			TOT
	R=1	R=0	*tot*	R=1	R=0	*tot*	
Female	3	2	*5*	3	2	*5*	10
Male	1	2	*3*	5	2	*7*	10
tot	4	4	*8*	8	4	*12*	20

Gender	R=1			R=0			TOT
	Y=1	Y=0	*tot*	Y=1	Y=0	*tot*	
Female	3	3	*6*	2	2	*4*	10
Male	1	5	*6*	2	2	*4*	10
tot	4	8	*12*	4	4	*8*	20

Commented [55]: تفاوت جدول sepration و sufficiency

Sufficiency example 1 / R=1

Gender	R=1			R=0			TOT
	Y=1	Y=0	*tot*	Y=1	Y=0	*tot*	
Female	3	3	*6*	2	2	*4*	10
Male	1	5	*6*	2	2	*4*	10
tot	4	8	*12*	4	4	*8*	20



$$\mathbb{P}\{Y=1 \mid R=1, A=f\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=1,A=f\}}{\mathbb{P}\{R=1,A=f\}} = \frac{\frac{3}{20}}{\frac{6}{20}} = 0.50$$

$$\mathbb{P}\{Y=1 \mid R=1, A=m\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=1,A=m\}}{\mathbb{P}\{R=1,A=m\}} = \frac{\frac{1}{20}}{\frac{6}{20}} = 0.17$$

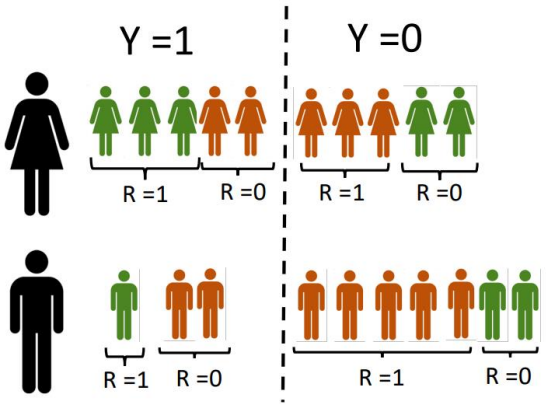
A={female *f*, male *m*}

Wrongly classified

Correctly classified

Sufficiency example 1 / R=0

Gender	R=1			R=0			TOT
	Y=1	Y=0	*tot*	Y=1	Y=0	*tot*	
Female	3	3	*6*	2	2	*4*	10
Male	1	5	*6*	2	2	*4*	10
tot	4	8	*12*	4	4	*8*	20



$$\mathbb{P}\{Y=1 \mid R=0, A=f\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=0,A=f\}}{\mathbb{P}\{R=0,A=f\}} = \frac{\frac{2}{20}}{\frac{4}{20}} = 0.50$$

$$\mathbb{P}\{Y=1 \mid R=0, A=m\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=0,A=m\}}{\mathbb{P}\{R=0,A=m\}} = \frac{\frac{2}{20}}{\frac{4}{20}} = 0.50$$

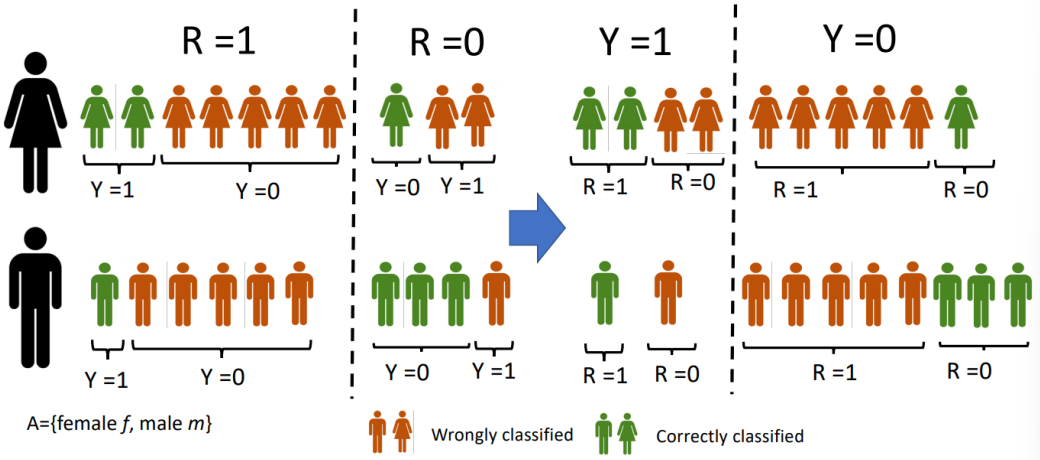
A={female *f*, male *m*}

Wrongly classified

Correctly classified

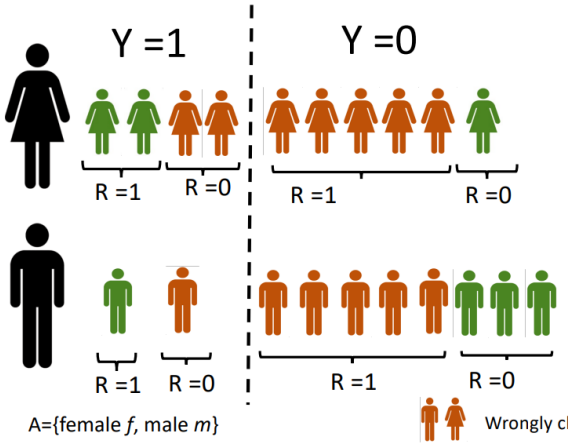
Separation – example 2 ---> Sufficiency example 2

Separation – example 2 → Sufficiency example 2



Sufficiency example 2 / R=1

Gender	R=1			R=0			TOT
	Y=1	Y=0	*tot*	Y=1	Y=0	*tot*	
Female	2	5	*7*	2	1	*3*	10
Male	1	5	*6*	1	3	*4*	10
tot	3	8	*13*	3	4	*7*	20



$$\mathbb{P}\{Y=1|\,R=1, A=f\}$$

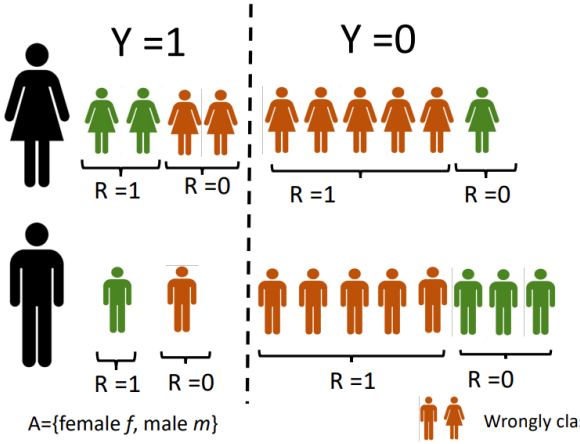
$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=1,A=f\}}{\mathbb{P}\{R=1,A=f\}} = \frac{\frac{2}{20}}{\frac{7}{20}} = 0.29$$

$$\mathbb{P}\{Y=1|\,R=1, A=m\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=1,A=m\}}{\mathbb{P}\{R=1,A=m\}} = \frac{\frac{1}{20}}{\frac{6}{20}} = 0.17$$

Sufficiency example 2 / R=0

Gender	R=1			R=0			TOT
	Y=1	Y=0	*tot*	Y=1	Y=0	*tot*	
Female	2	5	*7*	2	1	*3*	10
Male	1	5	*6*	1	3	*4*	10
tot	3	8	*13*	3	4	*7*	20



$$\mathbb{P}\{Y=1|\,R=0, A=f\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=0,A=f\}}{\mathbb{P}\{R=0,A=f\}} = \frac{\frac{2}{20}}{\frac{3}{20}} = 0.67$$

$$\mathbb{P}\{Y=1|\,R=0, A=m\}$$

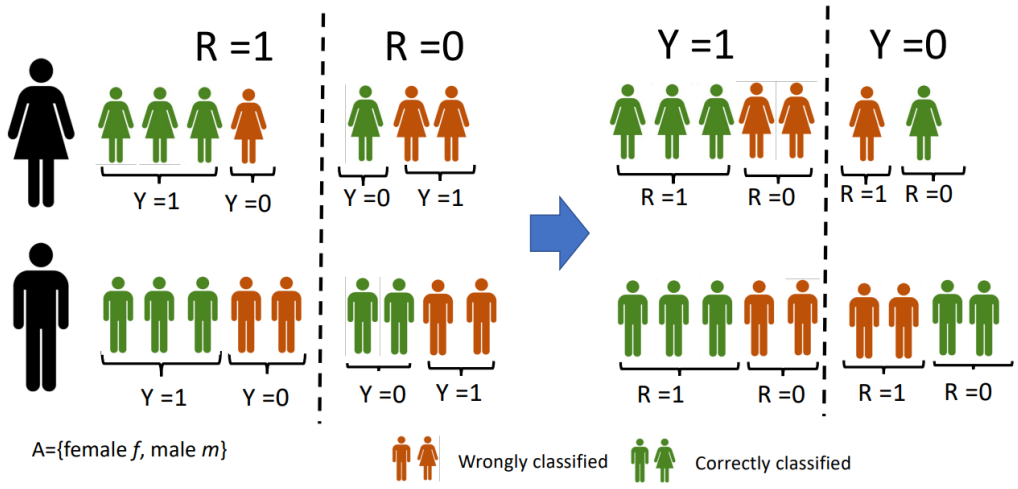
$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=0,A=m\}}{\mathbb{P}\{R=0,A=m\}} = \frac{\frac{1}{20}}{\frac{4}{20}} = 0.25$$

A={female *f*, male *m*}

Wrongly classified Correctly classified

Separation – example 3 —> Sufficiency example 3

Separation – example 3 → Sufficiency example 3

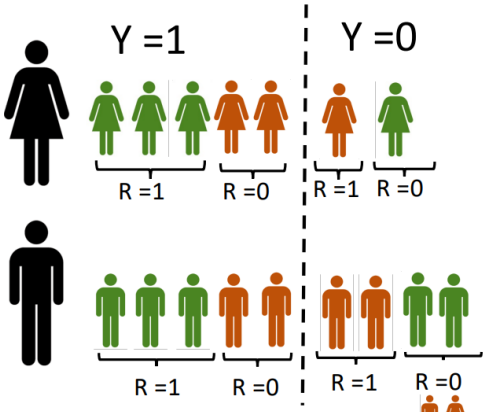


A={female *f*, male *m*}

Wrongly classified Correctly classified

Sufficiency example 3 / R=1

Gender	R=1			R=0			TOT
	Y=1	Y=0	*tot*	Y=1	Y=0	*tot*	
Female	3	1	*4*	2	1	*3*	7
Male	3	2	*5*	2	2	*4*	9
tot	6	3	*9*	4	3	*7*	16



A={female *f*, male *m*}

Wrongly classified

Correctly classified

$$\mathbb{P}\{Y=1|\,R=1,A=f\}$$

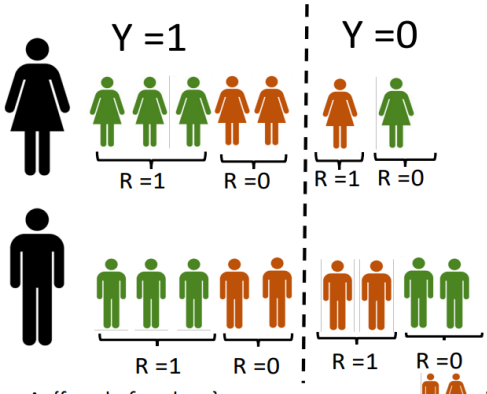
$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=1,A=f\}}{\mathbb{P}\{R=1,A=f\}}=\frac{\frac{3}{20}}{\frac{4}{20}}=0.75$$

$$\mathbb{P}\{Y=1|\,R=1,A=m\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=1,A=m\}}{\mathbb{P}\{R=1,A=m\}}=\frac{\frac{3}{20}}{\frac{5}{20}}=0.60$$

Sufficiency example 3 / R=0

Gender	R=1			R=0			TOT
	Y=1	Y=0	*tot*	Y=1	Y=0	*tot*	
Female	3	1	*4*	2	1	*3*	7
Male	3	2	*5*	2	2	*4*	9
tot	6	3	*9*	4	3	*7*	16



A={female *f*, male *m*}

Wrongly classified

Correctly classified

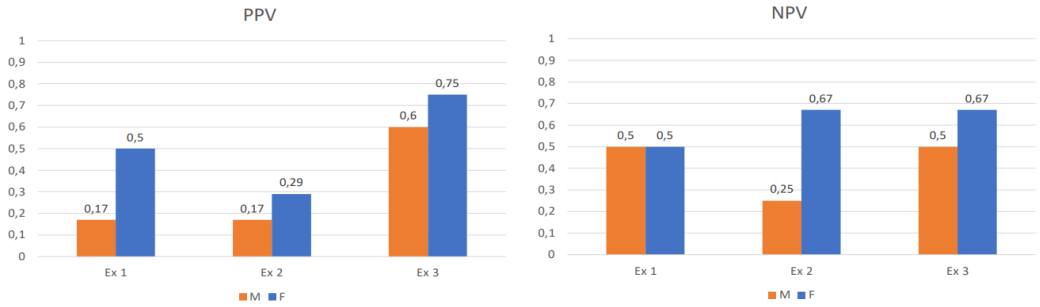
$$\mathbb{P}\{Y=1|\,R=0,A=f\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=0,A=f\}}{\mathbb{P}\{R=0,A=f\}}=\frac{\frac{2}{20}}{\frac{3}{20}}=0.67$$

$$\mathbb{P}\{Y=1|\,R=0,A=m\}$$

$$\frac{\mathbb{P}\{Y=1\}\cap\mathbb{P}\{R=0,A=m\}}{\mathbb{P}\{R=0,A=m\}}=\frac{\frac{2}{20}}{\frac{4}{20}}=0.50$$

Sufficiency - Summary of the 3 examples



Calibration and sufficiency

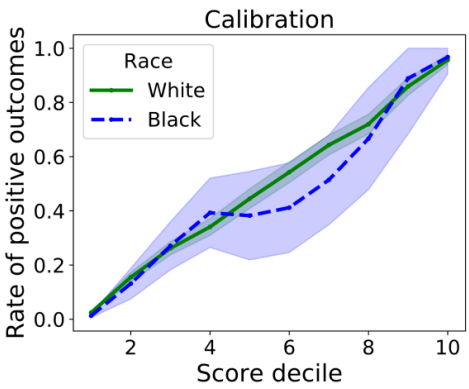
A score R is calibrated if $\mathbb{P}\{Y = 1 | R = r\} = r$

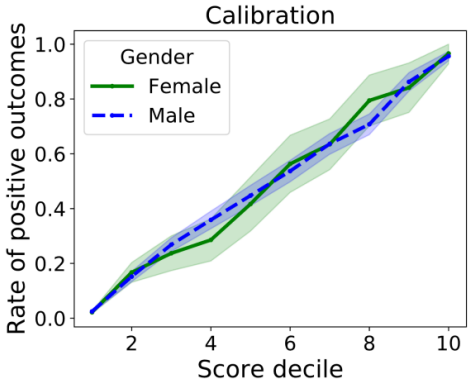
For all instances with an R score, there is a fraction r of positive instances. Formally, for each set S: $\mathbb{P}\{Y = 1 | R = r, X \in S\} = r$

Calibration is obtained from the latter equation:
$\mathbb{P}\{Y = 1 | R = r, \mathrm{A} \in \mathrm{a}\} = \mathrm{r}, \qquad \forall r, \forall a$

Commented [56]: just you need to know that if I say that the classifier is calibrated, means that sufficiency is respected.

Example





An example of very bad calibration is here. uh race so this algorithm is not well calibrated by ethnic group or at least by race considering only white and black because here we have multiple levels of the classification from one to 10. So we don't have only the binary case. And here we have the rate of positive outcomes. So these are the white, the green line and the

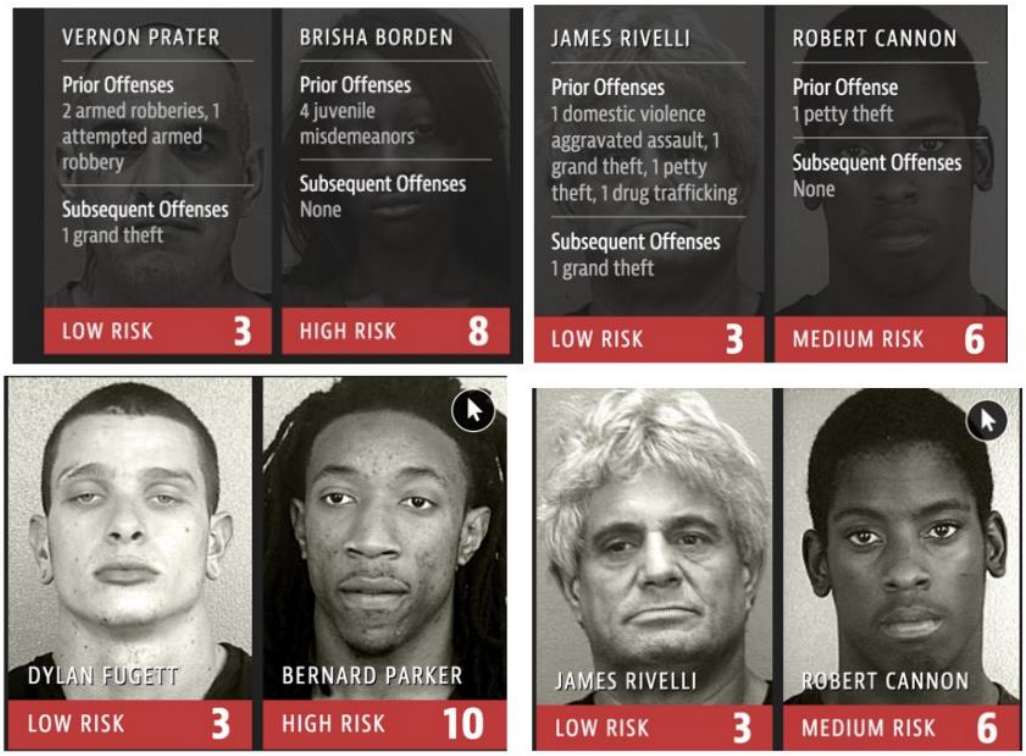
black is quite different for in this area. This is a calibration plot. So for every level of the classification. You ompute the rate of positive outcomes.

Chapter 06-Study case: COMPAS

The COMPAS system

The COMPAS* system is used in some U.S. court systems to estimate the probability of recidivism of a convicted offender.

*Correctional Offender Management Profiling for Alternative Sanctions



COMPAS: SUMMARY OF THE FAIRNESS CRITERIA

- Independence NOT satisfied:
  - Probability (Risk | African-american) = 0.58
  - Probability (Risk | Caucasian) = 0.33
- Separation NOT satisfied:
  - True positive rate, Caucasian: 0.50
  - False positive rate, Caucasian: 0.22

• True positive rate, African American: 0.72 • False positive rate, African American: 0.42

• Sufficieny satisfied (if relaxed):

• Positive predictive value, caucasian : 0.59 • FOR, caucasian: 0.29

• Positive predictive value, african american: 0.65 • FOR, african american: 0.35

Commented [57]: اگر tolerance level of 5-6 percent را در نظر بگیریم

Commented [58]: Here there is an error, it's not a negative predictive value, it's false submission rate

COMPAS: KEY FACTS (1/2)

• Pro Publica found that:

• White defendants who got rearrested were nearly twice as likely to be misclassified as low risk

• Black defendants who did not get rearrested were nearly twice as likely to be misclassified as higher risk

COMPAS: KEY FACTS (2/2)

• Equivant/Northpointe, the developers of COMPAS, replied:

• COMPAS satisfies calibration: scores have predictive values very similar by defendant's race. E.g., with score = 7, 60% of white defendants were re-arrested and 61% of black defendants were rearrested

• COMPAS satisfies equal positive predictive values: among those labeled higher risk, the proportion of defendants who got rearrested is approximately the same regardless of race (0.59 CA and 0.65 AA in our analysis).

COMPAS: SOME REFLECTIONS ON AGE How should we act ?

• Among the recidivists:

• 27% is 25 years or less

• 58% is between 25 and 45 years old

• 15% is more than 45 years old

• Among the low risk:

• 14% is 25 years or less

- 57% is between 25 and 45 years old
- 29% is more than 45 years old

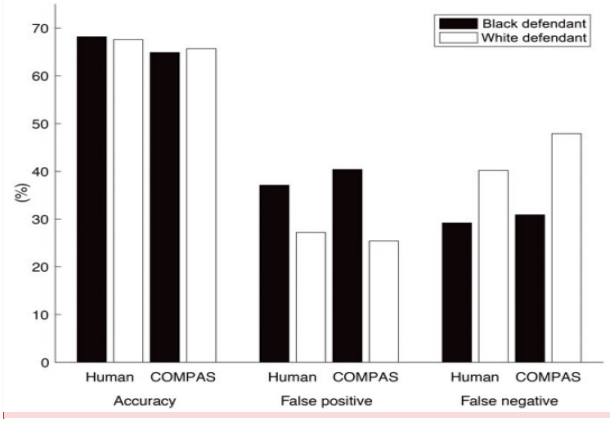
• Among the high risk:

- 34% is 25 years or less
- 59% is between 25 and 45 years old
- 7% is more than 45 years old

COMPAS: SOME REFLECTIONS

• A study by Prof. J. Dressel and Prof. H. Farid compared COMPAS predictions with those of a random sample of people with little or no expertise on criminal justice

• Result were very similar ->



COMPAS: SOME REFLECTIONS

• In the same study, it was proven that "despite COMPAS's collection of 137 features, the same accuracy can be achieved with a simple linear classifier with only two features"

- namely, age and prior criminal history—traditionally used to predict recidivism.

Commented [59]: Consider that there were some studies in which that showed that the accuracy of compass was not so better than the accuracy of using human judges.

Commented [60]: Regarding the false positives, the Compass case slightly outperformed the human, the human judgment for black defendants and vice versa. There was a slight improvement in white cases from the human judgment

Table 2 Algorithmic predictions from 7214 defendants.

Logistic regression with 7 features (A) (LR_7), logistic regression with 2 features (B) (LR_2), a nonlinear SVM with 7 features (C) (NL-SVM), and the commercial COMPAS software with 137 features (D) (COMPAS). The results in columns (A), (B), and (C) correspond to the average testing accuracy over 1000 random 80%/20% training/testing splits. The values in the square brackets correspond to the 95% bootstrapped [columns (A), (B), and (C)] and binomial [column (D)] confidence intervals.

	(A) LR_7	(B) LR_2	(C) NL-SVM	(D) COMPAS
Accuracy (overall)	66.6% [64.4, 68.9]	66.8% [64.3, 69.2]	65.2% [63.0, 67.2]	65.4% [64.3, 66.5]
Accuracy (black)	66.7% [63.6, 69.6]	66.7% [63.5, 69.2]	64.3% [61.1, 67.7]	63.8% [62.2, 65.4]
Accuracy (white)	66.0% [62.6, 69.6]	66.4% [62.6, 70.1]	65.3% [61.4, 69.0]	67.0% [65.1, 68.9]
False positive (black)	42.9% [37.7, 48.0]	45.6% [39.9, 51.1]	31.6% [26.4, 36.7]	44.8% [42.7, 46.9]
False positive (white)	25.3% [20.1, 30.2]	25.3% [20.6, 30.5]	20.5% [16.1, 25.0]	23.5% [20.7, 26.5]
False negative (black)	24.2% [20.1, 28.2]	21.6% [17.5, 25.9]	39.6% [34.2, 45.0]	28.0% [25.7, 30.3]
False negative (white)	47.3% [40.8, 54.0]	46.1% [40.0, 52.7]	56.6% [50.3, 63.5]	47.7% [45.2, 50.2]

Commented [61]: A(LR7) with 7 features
B(LR2) with 2 features
C(SVM)
D(COMPAS) with 137 features

That are case A, a logistic regression LR with only seven features. and B, is a logistic regression LR again with only two features. port vector machines with machine with seven features.
nonlinear super vector machine. And then the last term of comparison is the compass case with 137 features.

Commented [62]: we have, especially in Europe, less in the United States, but we have quite universally agreed, at least in terms of good practices, on data minimization. You should collect only the data that is strictly necessary for your... prediction goals.
هدف به حداقل رساندن داده ها و اینکه باید فقط داده هایی را جمع آوری کنید که برای اهداف پیش بینی شما کاملاً ضروری است.

the goal of this study was to understand whether the data minimization principle was used in Compass and what were the performances achievable by much simpler models.

هدف این بود ببینیم آیا اصل کمینه‌سازی داده‌ها در Compass مورد استفاده قرار می‌گیرد یا نه

the overall accuracy Row reached by a **linear aggression** with only 7 features instead of 137 case the is the same basically. 66% with a confidence interval of 95% around 64 and 68.

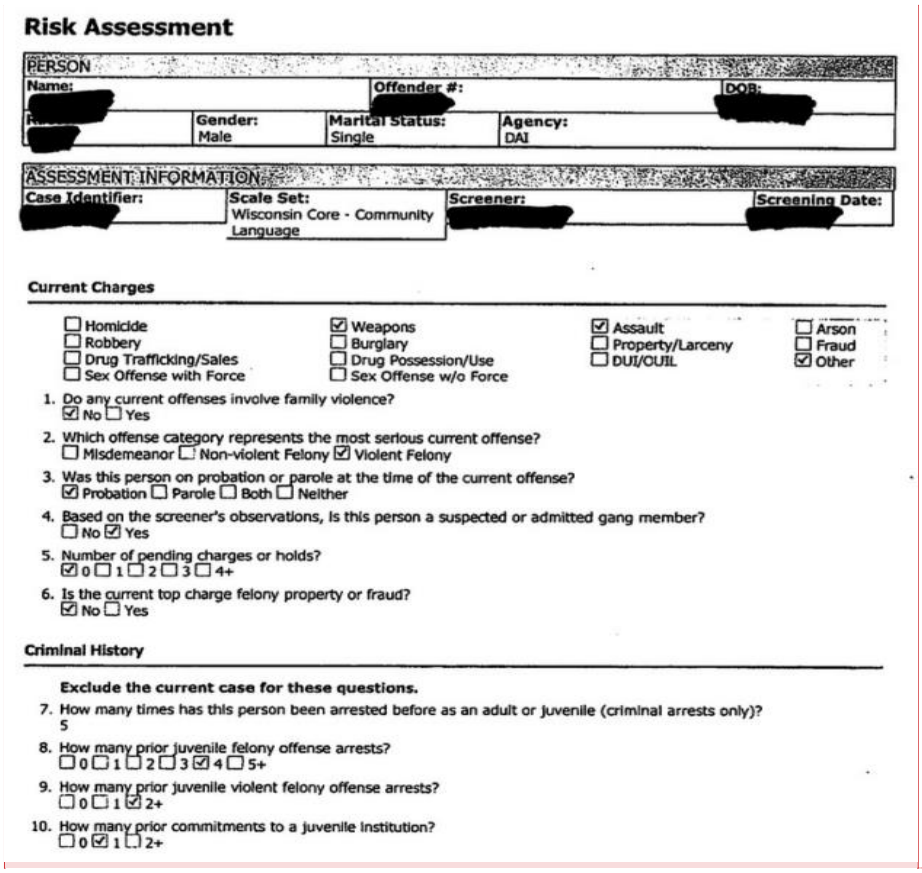
A logistic regression with only two features reached, again, the same accuracy level, 66%.

As you can see, 66% again and 60% for the logistic regression with seven features. Same level with two features. And about Compass. We have again similar values, which only accuracy for white is slightly higher, 67%, and slightly lower than previously for black people. If we look at the **false positive**, third and fourth row, for black. Again, we have the highest percentage was 45%, logistic regression with two features, but it's not so far away from the **44.8% of Compass**. This is a percentage of **false positive** for black. The best result was reached with support vector machine. Also in terms of **false positive** for white, 20%. Regarding the **false negative**, the support vector machine performed worst, compass 28%, and simpler logistic regression reached a minimum of 21%. For white, we had higher percentages of **false negatives**. And this is in line with the problem of compass, where **false negatives** were higher for whites. This is an advantage, because you are not predicted as a high risk why you are a recidivist. 47% for compass, and that is the same. basically as for the other logistic regression, a bit better than a sub-vector machine. Of course, I will never ask you about the percentages, but the takeaway is that you could achieve similar results or even better using a much less data. And why using a much less data? We will have now a closer look to the data sources of Compass to understand the why. it is important from the point of view of at least of human dignity to **use less data**.

Commented [63]: false positive for black: the best result was SVM
false negative for black:the Worse result was SVM

A closer look to the data sources

Commented [64]: این یک نمونه از فرمی است که سامل 137 ویژگی است که توسط پلیس و مجرم پر میشود
سوالی از قبیل
اطلاعات بسیار شخصی مثل نام و وضعیت تاهل و جنسیت
اطلاعاتی از محل زندگی طرف
اطلاعاتی ئر مورد اینکه قبلا مرتکب جرم شده یا نه(سابقه جنایی)
یک بخش در مورد سوء مصرف مواد
محل سکونت، ثبات زندگی
اوقات فراغت، تفریح، انزوای اجتماعی
اتهامات فعلی



Risk Assessment

PERSON

Name: [redacted]	Offender #: [redacted]	DOB: [redacted]
[redacted]	Gender: Male	Marital Status: Single	Agency: DAI

ASSESSMENT INFORMATION

Case Identifier: [redacted]	Scale Set: Wisconsin Core - Community Language	Screener: [redacted]	Screening Date: [redacted]

Current Charges

- ☐ Homicide
- ☐ Robbery
- ☐ Drug Trafficking/Sales
- ☐ Sex Offense with Force
- ☑ Weapons
- ☐ Burglary
- ☐ Drug Possession/Use
- ☐ Sex Offense w/o Force
- ☑ Assault
- ☐ Property/Larceny
- ☐ DUI/OUIL
- ☐ Arson
- ☐ Fraud
- ☑ Other

1. Do any current offenses involve family violence?
   ☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
   ☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
   ☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
   ☐ No ☑ Yes
5. Number of pending charges or holds?
   ☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
   ☑ No ☐ Yes

Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?
   5
8. How many prior juvenile felony offense arrests?
   ☐ 0 ☐ 1 ☐ 2 ☐ 3 ☑ 4 ☐ 5+
9. How many prior juvenile violent felony offense arrests?
   ☐ 0 ☐ 1 ☑ 2+
10. How many prior commitments to a juvenile institution?
    ☐ 0 ☑ 1 ☐ 2+

یکسری بخش های این فرم هم از قبل توسط پلیس پر میشود مثل اینکه
این جنایت خشن بوده یانه / اسلحه داشته یا نه/ این فرد تابه حال آزادی مشروط داشته یا نه/عضو باندی بوده یا نه/قبلا چند بار دستگیر شده/چند تا از جرایمش جرایم نوجوانی بوده/چند تاش خشونت توش بوده/
باز یه سری سوال از شخص پرسیده میشه
چه کسی شما رو بزرگ کرده فقط پدر هر دو یا.../و سوالاتی در مورد جرایمی ک انجام داده تا در ابتدا دلایل انجام رو پیدا کنن

This is **very different** from taking this information, turn it into a variable, and make a prediction with this and other information. Because if this aspect is statistical, will be a statistical significant predictor, it means that an individual with this kind of personal characteristics for instance, raised by adoptive parents. Let's say that this will be this level of the variable will be significant. But it also means that whenever you have you are these characteristics, you are raised by adoptive parents, you are somehow more probable to receive a high risk.

there is a huge difference when you turn information that should be used in a qualitative way into a rough statistical pattern.

درک کلی محیط اجتماعی که یک فرد در آن بزرگ شده است، که در آن ممکن است در زندگی با مشکلاتی روبرو شود، مهم است، اما اگر یکی از این سطوح عامل مهمی برای پیش بینی باشد، به راحتی می تواند به یک تبعیض خودکار تبدیل شود.

It's important to understand the whole social environment in which a person is raised, in which he or she could encounter difficulties in life, but can easily become an automatic discrimination if one of these levels will be an important factor for the prediction.

در مورد تمام سوالاتی که پرسیده میشود ممکن است مجرم واقعیت را نگوید. پس باید این را در نظر داشت مثلا ممکن است عضو یک باند باشد اما بگوید نه نیست. یا مثلا آیا در زمان دستگیری الکل مصرف می کردید؟ در این مورد بگه، بله.

اول، ما تحت یک زمینه قانونی هستیم که در آن شما باید حقیقت را بگویید. هر زمینه حقوقی ایجاب می کند که حقیقت را بگویید، اما... تنظیمات حقوقی ایالات متحده و همچنین تنظیمات حقوقی بریتانیا یک جنبه مهم، یک ارتباط مهم در این جنبه قرار می دهند. اگر هر یک از شما برای گرفتن مجوز ورود به سمت ایالات متحده پرواز کرد، باید اعلام کنید که تروریست نیستید. این ممکن است عجیب به نظر برسد. اول از همه، اجباری است. ثانیاً، مهم است زیرا در آن زمینه قانونی، شما مسئولیت آن ادعا را بر عهده می گیرید. بنابراین اگر تروریست هستید، اگر تروریست هستید و اعلام می کنید که نیستید، اتهام شما بسیار بیشتر خواهد بود.

But remember, this variable then turns, this information then is turned into a variable. And if it is statistical significant, you are building a **risk score based on**, even on this kind of information.

this information is then turned into variables. We have seen that there are reliability issues. And last but not least, you get **discrimination issues** because statistical patterns may inform a prediction based on personal characteristics.

COMPAS: summary of issues

- **Intervention vs prediction:** high accuracy won't solve the problem
- **Target-construct mismatch:** re-arrest != re-offense/recidivism (measurement bias)
- **Distribution shifts:** temporal + geographic (evaluation bias)
- **Limits to prediction:** comparable to linear model with 2 features
- **Disparate performance**: lack of independence and separation
- **Lack of contestability**: subjects cannot challenge the information
- **Goodhart's law:** social desirability bias on

نتیجه گیری

Let's close the Compass case just with a summary of the issues. Using a schema that you can find here, where each of these aspects are explained. **First of all**, Intervention versus prediction. Even if you have a high, you can get a high accuracy or predictions, and this was not the case in Compass, at least for certain subgroups, you are not solving the problem of recidivism. **You are just predicting the probability of being rearranged**. And this is linked with the second aspect, target construct mismatch. **You have a construct and this is a measurement bias**. You don't on the target variable because you are not really measuring a recidivism. **You are measuring the probability of being a rearrested**. It is different from recidivism. We might have in Compass ahead distribution shift that is evaluation bias because of using data from a specific range of time and then keep using the model for a while and also for geographical differences in the criminality. We have seen that Compass was really comparable to linear models. The performances were comparable to linear models with only two features or with seven features. We have seen that there was a lack of independence and separation. **The model was calibrated**, **but did not satisfy independence and separation**. Then **the subjects could not contest, could not challenge and contest the information and even the decision.** They were not aware even of a risk score on them. And then, good at throw, what I was just mentioning earlier, you might, as a defendant, be pushed to change your answers to the questions because of several factors. Because you don't want to be in trouble after you are out of jail, because you would like to influence the decision of the judge, et cetera, et cetera. You could never know what is in the mind of a person under judgment when answering such personal questions. So you might lie on your personal history. in a way or in the other in order to influence the judge.

we see the **reflexivity problem**, the canvas law. Whenever an indicator is used for an evaluation of a person, a group of persons will be subject to social pressure. Social corruption is also called in technical terms. henever your measure becomes a target, then it ceases to be a good measure. هر گاه معیار(اندازه گیری)شما به یک هدف تبدیل شود، دیگر معیار خوبی نیست.

Commented [65]: مداخله در مقابل پیش بینی

Commented [66]: شما بر روی متغیر هدف نیستید زیرا واقعاً تکرار جرم را اندازه نمی‌گیرید

Commented [67]: هرگاه از شاخصی برای ارزیابی یک فرد استفاده شود، گروهی از افراد تحت فشار اجتماعی قرار می گیرند.

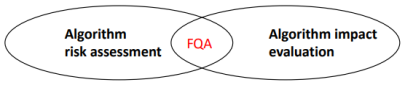
Commented [68]: فساد اجتماعی:شما سعی می کنید رفتار خود را تغییر دهید تا آن شاخصی را که علیه شما استفاده می شود بهینه کنید. مثل مثال انتخاب برای دانشگاه داشتن تعداد x تا مقاله و کشتن مارهای کبری در هند

Chapter 07- Fairness Qualitative Assessment

Fairness qualitative assessment

-What

- Identify issues of fairness



- When

- Design phase, but also during and after deployment

- Who by ?

- Creators and/or commissioners of the algorithmic system

- Researchers, policymakers

- Students

- Possibly with the help of impacted stakeholders

Fairness qualitative assessment

1) What **social conflict(s)** arise from the design/operation of a system?

2) Which stakeholders are adversely affected by this conflict? How?

3) What values/social interests are at stake in this case?

4) How could the value/interest conflict be (fairly) resolved?

Example

http://saferroute.appspot.com/heatmap

An app, basically a service, digital service that was designed with data from criminal records of the city of Edinburgh. building a recommendation for itinerary, urban itineraries in order to minimize the probability to incur into a crime. if you want a safe route, you use this recommendation algorithm that is built on top of historical crime data.

Commented [69]: FQA:
it is in the middle between algorithm risk assessment or algorithm impact evaluation. The difference is that you make a risk assessment before you deploy your system in design phase. You make an impact evaluation usually after or when your system is under testing, under deployment.

You can use the FQA in both cases. The goal is to identify issues of fairness, in both design phase during after deployment.

1. What social conflict(s) arises from the design/operation of a system?

Taylor-Russell (T-R) diagram

a.k.a. "Confusion matrix"

		Prediction	
		Negative	*Positive*
Event	*Present*	False negative	True positive
	Absent	True negative	False positive

		Prediction	
		Low risk of assault	**High risk of assault**
Event	**Criminal intention or circumstance**	Dangerous area/circumstance misclassified as safe	Danger avoided
	No assault	Safe area/circumstance correctly labelled	Safe area/circumstance misclassified as dangerous

the **event** is, and this is difficult to compute, this is an intrinsic limitation of such kind of social outcome prediction, that is, there was a crime or not, or there were the circumstances for a crime or not. or there was no assault. So there is some practical difficulties in tracking whether the event occurred or not. so whenever you have a low risk of assault as a prediction, and instead that you had a crime happening, that is that a dangerous area or circumstance was misclassified as safe (FN). **This is an error**. The **other error** is whenever you have a high risk of assault predictor, but nothing happened. So you made a longer route, but it was not necessary.So safe areas misclassified as dangerous (FP).

Commented [70]: the social conflicts arise in occurrence of the errors.

2. Which stakeholders are adversely affected by this conflict? How?

Commented [71]: چه کسانی ذینفعانی هستند که تحت تأثیر منفی این درگیری ها قرار می گیرند

Stakeholders' identifications

- Direct stakeholder (or ”user”)
  - persons who interact directly with a tool or technology
- Indirect stakeholders
  - persons who do not use the technology under consideration
  - however, they are impacted by it

Stakeholders - safer route

- Direct stakeholders:
  - The users of the app
    - especially those who are new in the city/neighborhood, e.g., tourists, new residents, etc.
      - woman may use the app more often
  - (The investors/owners of the app)
- Indirect stakeholders
  - Residents living in the areas covered by the app
  - Business in the areas covered by the app
  - (Police corps of the involved areas)

2. Which stakeholders are adversely affected by this conflict? How? – Safer route

- False positives
  - Business owners: loss of customers, reputation damage
  - Residents: loss economical values for properties, reputation damage
- False negatives
  - The users of the app suffer assaults that might otherwise not have happened.

		Prediction	
		Low risk of assault	High risk of assault
Event	Criminal intention or circumstance	Dangerous area/circumstance misclassified as safe	Danger avoided
	No assault	Safe area/circumstance correctly labelled	Safe area/circumstance misclassified as dangerous

Please notice that if you do this analysis at design phase, you can try to counteract these negative impacts. But if you do these analysis only after deployment, then you need to wait at least a few years. It is very difficult that a system that predicts social outcomes will be able to have 0% of errors in false positive and false negatives is impossible. You will always have some kind of impact.

Commented [72]: ذینفعان مستقیم، SeferRoute در مورد کاربران برنامه هستند. یا افرادی که دانش عمیقی از یک شهر، یک منطقه شهری خاص دارند، اصلاً به این نوع برنامه نیاز ندارند و گردشگران، افرادی که در شهر یا محله خاصی تازه کار هستند، ساکنان جدید و غیره

Commented [73]: ساکنانی هستند که در آن شهرزندگی می کنند، یا حداقل در مناطقی که توسط برنامه به عنوان خطرناک یا غیر خطرناک رتبه بندی شده اند. همچنین کسانی که فعالیت های تجاری در آن مناطق انجام میدهند.

Commented [74]: (FP)دایره ای که گفته شده دنجرس):اگر به جدولی که قبلاً پر شد نگاه کنیم، می توانیم بگوییم که نتایج مثبت کاذب ممکن است بیشتر بر صاحبان مشاغل تأثیر بگذارد، صاحبان مشاغل واقع در مناطقی که به اشتباه به عنوان خطرناک طبقه بندی شده اند، زیرا ممکن است افراد کمتری به آنجا مراجعه کنند. یا ساکنان، اگر افراد بسیار زیادی از این برنامه استفاده کنند، ممکن است ملک شما در مورد محله خاصی شهرت بدی داشته باشد و ملک شما ممکن است ارزش اقتصادی خود را از دست بدهد زیرا ممکن است تحت تأثیر قرار بگیرند.
(FN)دایره ای که گفته شده(Safe: ذینفعانی که بیشتر تحت تاثیر منفی های کاذب قرار می گیرند، کاربران اپلیکیشن هستند. زیرا ممکن است از منطقه ای عبور کنید، زیرا الگوریتم توصیه می کند، که در واقع خطرناک است، حتی اگر سوابق کیفری کافی وجود نداشته باشد، اما در واقع خطرناک است. خوب، در نظر بگیرید که اجازه دهید فقط جنایات خشونت آمیز را در نظر نگیریم.

3. What values/social interests are at stake in this case?

Values identification

• Values

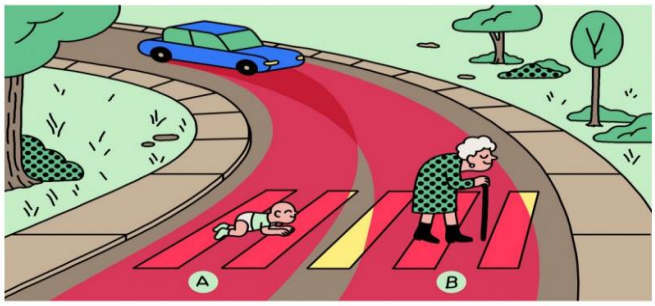
• "the principles or standards of a person or society, the personal or societal judgement of what is valuable and important in life" (Oxford Dictionary) à what is important to people in their lives

• often related to ethics and morality (but not necessarily!)

• hierarchy of values constitute what we are/want to be

• impossible to avoid conflicting values between individuals and social groups

Here we are principally interested in values that concern:
- **human welfare (e.g., education, health, income)**
- **justice (attributive or distributive)**

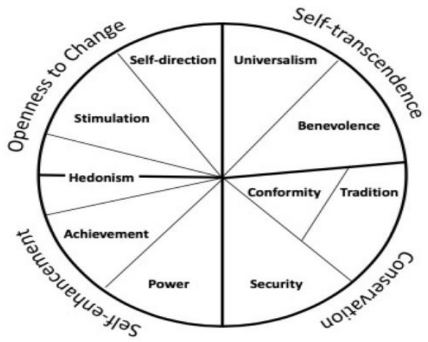


http://moralmachine.mit.edu/

You will be. asked to decide on 13 scenarios in which you have an autonomous vehicle, vehicle that is in such a situation in which you can only decide whether the vehicle should go straight, so take no action and kill either some people on the street or the people inside the car or take an action. In order to avoid this child, you take action. So the car that is autonomous should decide to make a turn, and the result, the consequence, will be killing a woman, an older lady. So you will be asked 13 times questions that are not true.

Commented [75]: یک مثال است که برای امتحان هم مهم نیست
که از شما خواسته شد در مورد MIT این یک تست از دانشگاه 13 سناریو تصمیم گیری کنید
به عنوان مثال، مشخص شد که در جوامع غربی، اکثر افرادی که در این نوع آزمون شرکت می‌کردند، از امتیاز برخوردار بودند. با احترام به جوان ترین افراد پس سعی کنید از کشتن کودک خودداری کنید. پس نجات بچه و کشتن خانم، پیرزن. در جوامع شرقی، جوامع شرقی مانند چین، ژاپن، هند برعکس بود. معمولاً افرادی که در این آزمایش شرکت می‌کردند، پیرزن را نجات می‌دادند و نوزاد خردسال یا افراد جوان‌تر را می‌کشتند. این به این دلیل است که انواع متفاوتی از جامعه است که حول ارزش ارزش های مختلف مرتبط تر با افراد مسن، ارزش اجتماعی بالای افراد مسن ساخته شده است. این فقط نمونه‌ای از جنبه‌های مختلف فرهنگی و ارزش‌های مختلف بود که فرهنگ، هنجارها، قوانین را از نظر قوانین حقوقی جوامع به‌طور متفاوتی شکل می‌دهند.

Schwartz's basic values



Picture from Dignum, V. (2019)

Commented [76]: لیست بی نهایتی از ارزش ها وجود دارد و این فیلسوف شوارتز ارزش ها را در چهار دسته کلان دسته بندی کرد.
Openness to change is you give importance to self-direct your own life, to get stimuli in your life, to make fun in your life, to be better, to build your own skills, power. You want to get power, political power, economical power. You want to achieve that. Conservation, you feel more, you value more security in your life. Or you value a lot of tradition, conformity to a given set of rules, norms. And then self-transcendence. Attention

More specific examples of values

• Trust: "Refers to expectations that exist between people who can experience good will, extend good will toward others, feel vulnerable, and experience betrayal"

• Autonomy: "Refers to people's ability to decide, plan, and act in ways that they believe will help them to achieve their goals"

• Universal usability: "Refers to making all people successful users of information technology"

• Informed consent: "Refers to garnering people's agreement, encompassing criteria of disclosure and comprehension (for "informed") and voluntariness, competence, and agreement (for "consent") "

Values: further examples

• Privacy

  • the right to determine what information about yourself can be communicated to others

• Physical safety

  • absence of harm or injury

• Environmental sustainability

  • satisfying life needs without putting unnecessary strain on the earth's ecosystems

• Financial stability

• having stable income that allow safe housing, healthy food, and other biological and social necessities

• Free time

• possibility to regularly spend time away from work to enjoy other aspects of life

Values & software

• Explicitly supported

• can be in the form of design constraints or even formal requirements

• Designer values

• personal or professional values that a designer/developer/software architect brings into the design of a tool;

• not necessarily aligned with/supported by an explicitly supported value

• Stakeholder values

• What matters for specific stakeholders / stakeholders' groups

• They might by explicitly asked or elicited indirectly (e.g., values scenarios)

3. What values/social interests are at stake in this case? – safer route

• Economic and reputational security

• Residents/business owners of areas labelled as dangerous

• Personal safety

• Women, especially young

• Tourists or new residents

		Prediction	
		Low risk of assault	**High risk of assault**
Event	**Criminal intention or circumstance**	Dangerous area/circumstance misclassified as safe	Danger avoided
	No assault	Safe area/circumstance correctly labelled	Safe area/circumstance misclassified as dangerous

whenever we have the **false Positive**, we are **impacting economic security and the reputational value**, let's say, of residents, of commercial activities, people owning those commercial activities. So **these are the interests or values** that are impacted when you are negatively impacting these type of stakeholders.

In **false negative**, you are **impacting the personal safety**, this value, as we say, the particular stakeholders that you identified, new residents, tourists, women, et cetera.

4. How could the value/interest conflicts be (fairly) resolved?

Commented [77]: A startup, for instance, in Torino here, is working the field of news. collection of news from several newspapers and then presenting to you the most relevant news so that you don't have to pay yearly subscriptions to all newspapers, but you only pay a general subscription to Categorio Sport, for instance. An explicit value that was supported and translated into design requirements by this startup was to put a limit in the number of recommendations. In order to limit the time that you spend on your phone. This is an example of value that you can pursue and with a software and translate directly into software requirement. As an individual designer, you might also embed some value in your design.

Commented [78]: As an individual designer, you might also embed some value in your design. You want a website that is particularly well-looking and so this is a value for you that is nice looking. and you can translate it into technical requirements.

Commented [79]: we might have values from all the stakeholders that might be elicited from stakeholders and then translate into software requirements.

Commented [80]: whenever we have the false negatives, we are impacting economic security and the reputational value, reputational capital, let's say, of residents, of commercial activities, people owning those commercial activities.

How to solve value tensions?

- Examples:
- Solicit public feedback since design phase
    - Citizen jury, citizen assembly, deliberative polling, consensus conferences
- Enable software errors reports and reconfigure software during operation
- Value-based prioritization:
    - The need for safety is considered of greatest value and errors minimized accordingly
    - The economical value of property is considered of greatest value and errors minimized accordingly

First of all, you can try to **understand what the people want**, because this is an app that is for the people, is used to increase the security of the people, is a **first goal**. And then there are different ways of understanding what matters more for a given social group, for a given population. It is a jury, it is an assembly. Jury is when you vote. You ask selected people from different social groups, select representatives from different social groups to express their own interest. You might make conferences with citizens. You might make a polling, a questionnaire. You might so-called make consensus conferences, a series of conferences. When you describe the problem, you... you describe several perspectives and the people then discuss the whole citizenship, the whole population can at least potentially can intervene.

So these are just samples of actions that can be taken to mitigate or solve the value tensions that can arise from the execution of your system. As you can see, **this puts a different light on the issue of fairness**. It is not only a matter of fairness. of equilibrium of parity between those positive probability of being predicted, etc. **It's classifications and predictions of social outcomes.** impact more deeper aspects that need a **more qualitative reflections**, considerations, and then actions. If you take actions only on the base of statistical computation, so you will be in troubles. Also because, as I already said, you can optimize for one criteria of fairness, but not for the others. So you need to... understand that you need to use complementary tools and complementary points of view. This is a very simple example of running a qualitative analysis for fairness.

Commented [81]: You might say, okay, I start running the software, error reports from the people. So the people can start putting their feedback directly on the app and so that I can understand what's going on, what is their feedback, how they're using the app, and trying during operation to retune the application and try to solve, to mitigate the social tensions.

Commented [82]: یا می توانم بر اساس ارزش ها اولویت بندی کنم

Commented [83]: من می گویم نه این برنامه با هدف افزایش ایمنی ساخته شده است و من سعی می کنم ایمنی را بدون در نظر گرفتن منفی کاذب به حداکثر برسانم، حتی اگر 80 درصد آنها باشند، اهمیتی نمی دهم، بنابراین این یک رویکرد است.

Commented [84]: یا ممکن است بگوییم، نه، من می خواهم جنبه های اقتصادی، ارزش اقتصادی املاک و فعالیت های تجاری را حفظ کنم. و بنابراین می‌توانم حتی از نظر فنی سیستم طبقه‌بندی بهینه‌شده‌ام را برای این هدف کالیبره کنم.

Commented [85]: پس باید درک کنید که باید از ابزارهای مکمل و دیدگاه های مکمل استفاده کنید.

Chapter 08- Algorithmic Profiling of Job Seekers in Austria

THE CASE

Commented [ZK86]: کیس امتحان

- Algorithmic profiling system of job seekers
- Announced in October 2018, by the Public Employment Service in Austria (AMS)
    - profiling of job seekers independently started since before 2010
- Actual deployment of the system from July 2020
- Under legal disputation for discrimination

What are the main general characteristics of the system used by the public employment service in Austria, in order to profile job Seekers and decide which of them should receive assistance from the state.

ویژگی های اصلی سیستمی که توسط خدمات استخدامی دولتی در اتریش به منظور تعیین مشخصات جویندگان کار و تصمیم گیری در مورد اینکه کدام یک از آنها باید از دولت کمک دریافت کنند، مورد استفاده قرار گرفت.

to find a new job to receive personalized applications for instance. or funding for strengthening the curriculum with some courses attending some classes for example.

برای مثال شغل جدیدی برای دریافت برنامه های شخصی سازی شده پیدا کنید. یا بودجه برای تقویت برنامه درسی با برخی از دوره های حضور در برخی کلاس ها به عنوان مثال.

the deployment of the system was started in 2020. at the moment it is under legal disputation for issues of discrimination.

استقرار این سیستم در سال 2020 آغاز شد. در حال حاضر در مورد مسائل تبعیض تحت بحث قانونی است.

Political goal

it has a direct impact on the technical goal

- High-level political goal: make funding instruments more “effective”

more effective: means using the money in a better way. what is good money spent or not and what is public money that is not well spent on this specific sector or welfare that is providing assistance for finding a job.

• Support for reintegration into the job market dependent on the profile of the applicant

• AMS justification: according to AMS "experience", expensive active labor market programs do not significantly increase the chances of hiring both for high and low job seekers prospects

The decision of the employment service or employment agency was to find a way to identify the so-called **middle prospects** so people that are neither too good in their skills nor too bad, but people that are in this medium level, the **return on the investment** called lactic is **maximized**. because they have better chances to use the money that they receive and to find a job in a quite short time. This is the **justification** for the political goal that is behind the deployment of this system.

e.g: Phd , manager with high educational level doesn't need the help of government

VS low skills that it is very hard for them even if they get assistance to to find job

The AMS-Algorithm

• Classification of job seekers into three categories

  • Group H: High chance to find a job within 6 months

  • Group L: bad outlook of employment in the next 2 years

  • Group M: neither H nor L, as "mediocre" prospects → FOCUS OF FUNDING

• Regression model supposed to be used to assign individual scores to jobseekers and label each of them in on of the three categories

the political goal can be even a business goal also affected by economic needs to spend less money

Data-driven approach

• As all ML/AI algorithms, AMS algorithm works with historical data.

• It uses two types of data:

  • Data from the **procedure of registration to the AMS network**

  • Data from the **Main Association of Austrian Social Security Institutions, that collects personal data on the individuals** (e.g., gender, nationality, age).

We have a first problem here, the problem is that according to the GDPR if you collect the information, you should collect for a given purpose, and then if you change the purpose, you should ask again for permission or you should collect from scratch the data. so this is the very first problem that you are using data that was collected for other purposes.

Commented [ZK87]: اگر GDPR مشکل این است که طبق اطلاعات را جمع آوری کنید، باید برای یک هدف معین جمع آوری کنید، و سپس اگر هدف را تغییر دهید، باید دوباره مجوز بخواهید یا باید از ابتدا اطلاعات را جمع آوری کنید. داده ها. بنابراین این اولین مشکلی است که شما از داده هایی استفاده می کنید که برای مقاصد دیگر جمع آوری شده اند

„DATA CONSTELLATIONS“ recognizable pattern

Variables	Characteristics
Gender	M / F
Age Group	< 30 / 30 – 49 / 50+
Citizenship	Austria / EU / Non-EU Countries
Education	Grade school / Apprenticeship, vocational school / high- or secondary school, university
Health Impairment	Yes / No
Obligations of Care (only women)	Yes / No
Occupational Group	Production / Service
Regional Labor Market	Five categories (Type 1 – 5) for employment prospects in assigned job center
Prior occupational career	< 75% / >75% days of gainful employment within 4 years
Cases within 4 Year Intervals	0 cases / 1 case / 2 cases (1/year) / 3+ cases
Duation of Cases	0 cases > 6 months / 1+ cases > 6 months
Measures Claimed	0 / min. 1 supportive / min. 1 educational / min. 1 subsidized employment
Duration of Unemployment	Start / 3 / 6 / 9 / 12 / 15 / 18 / 21 / 24 / 30 / 36 / 48+ months

„Quality“ of AMS algorithm in test phase

- Precision [TP / (TP + FP)] as Performance Indicator
    - precision rates are only known for high and low segments
        - group H, precision in range of 80%-84%
        - group L, precision in range 81%-91%.

        → approx. 120,000 people with wrong results
- Limits of precision:
    - only shows the fraction of false positives
    - no information about **false negatives**.

•Discrepancies in error rates

- errors not distributed equally across populations
- e.g. between populations / constellations
- minorities particularly affected

Historical biases AMS reflects the high degree of historical inequality in the labor market
Examples:

- women at the beginning of unemployment approx. twice as often as men in the low segment
- 'migration background' and people with age > 50 systematically get lower scores
- "regional labor market" = proxy for social class
- in general, marginalized groups were disadvantaged in the classification

Other types of biases in AMS ALGORITHM

- Measurement bias:
  - weak abstraction of variables, hard thresholds
  - e.g., "care obligations" only apply to women. —> this is an error
- Aggregation bias:
  - inhomogeneity of chances within constellations
- Omitted variables bias and representation bias, due to:
  - regular changes in the labor market
  - change of social values, e.g. third gender
  - extraordinary events, e.g. recession 2008, COVID-19
  - legislative changes and local changes (e.g. the bankruptcy of a large company)

Values embedded in the software design

- Biography and skills reduced to variables of a prediction model
- Employability tied to the individual

Commented [88]: Aggregation bias is because we have even within the constellations, we have a particular groups that have higher employability than others.

Commented [89]: Your biography itself is reduced to a feature of a model. Qualitative aspects are lost. You don't have anymore prioritization on an individual analysis made by a person. It is not scalable, okay. But at least it can... can identify specific characteristics of the biography of the person that only a person can appreciate, can better qualify. If you turn it into a mere prediction model, you lose all of these aspects.

Commented [90]: you tie the responsibility of finding a job merely on the skills and the attributes and the characteristics. some of them even physical of the people, of the single individual. So whatever has been claimed as objective cannot be accepted as objective. Even the way the system has been designed, the goal of the system and given also the impact.

• Claims of objectivity versus value-laden selection of variables, categories, thresholds and performance indicators

AMS algorithm and organizational goals

Objectives as stated by the AMS:

1. Increase the efficiency of the counselling process

    *Risk of routine adoption of scores counteracts service orientation of agency

2. Increase the effectiveness of the use of funds

    *Coarse profiling of clients counteracts goal of effective use of resources

3. Standardize the granting of funding (vs. arbitrariness)

    *Disparate impact in the system‘s design

Other issues

• Lack of transparency about data collection and design details

• Lack of ability to verify, contest and remedy AMS decisions

• The use of sensitive information, initially collected for other purposes

• Social stigma and psychological consequences

if you don't get assistance for finding a job and you can not finding a job because of that your life becomes very difficult so you have long-term social and psychological consequences on the persons

AMS: summary of issues

• Intervention vs prediction: organizational goals not met

• Target-construct mismatch: mediocre prospect != not employed

• Distribution shifts: constellations, changes in law/market/values

• Limits to prediction: average precision around 80%

• Disparate performance: women, migrants, age 50+, minorities

• Lack of contestability: citizens cannot challenge the classification

• Goodhart's law: -

Commented [91]: We have seen that there was a quite broad profiling of people, and so there are doubts also on the fact that the funds were used in a quite effective way.

Commented [92]: the disparate impact that was measured is a problem because of discrimination. and because that you cannot claim to be objective when using objective criteria when giving the funding because you have then a disparate impact.

Commented [93]: we had other issues that are common when predictive optimization is used in the public, often not only in private companies, but also in the public administrations. The tests that have accountability duties that are higher, even higher than in the private sector. No transparency about data collection and design details.

Chapter 09- Detecting social welfare frauds in the Netherlands

SyRI : System Risk Indicator

- Organizational goal: support officials in investigating welfare frauds1
- System goal: predict individuals risks of welfare frauds
- Expected benefits:
    - Reduced inspection time and related cost savings
    - Better resource allocation
    - Less waiting time and administrative burdens
- Incorporated in Dutch legislation from 2014, deployed by Employee Insurance Agency and the Tax and Customs Administration, developed by the Totta Data Lab startup

Commented [94]: The organizational goal of the system was to support officials in, state officials in investigating when a welfare fraud was occurring.

Commented [95]: Thesystem goal was to predict the individual's risk. of welfare frauds, so getting a risk score for a specific person.

System Risk Indicator (SyRi)

3.2.6 SyRi (Systeem Risico Indicatie), the Netherlands

Country	The Netherlands
AI typology	Predictive Analytics, Simulation and Data Visualisation
Level of administration	Central/ Municipal
Policy Sector	Social Protection
Purpose	Enforcement
Main enablers	Sharing of data/resources, high data quality, Political leadership
(Expected) Impact	Improved inspection capabilities, improved welfare of children, reduced misuse of public funds

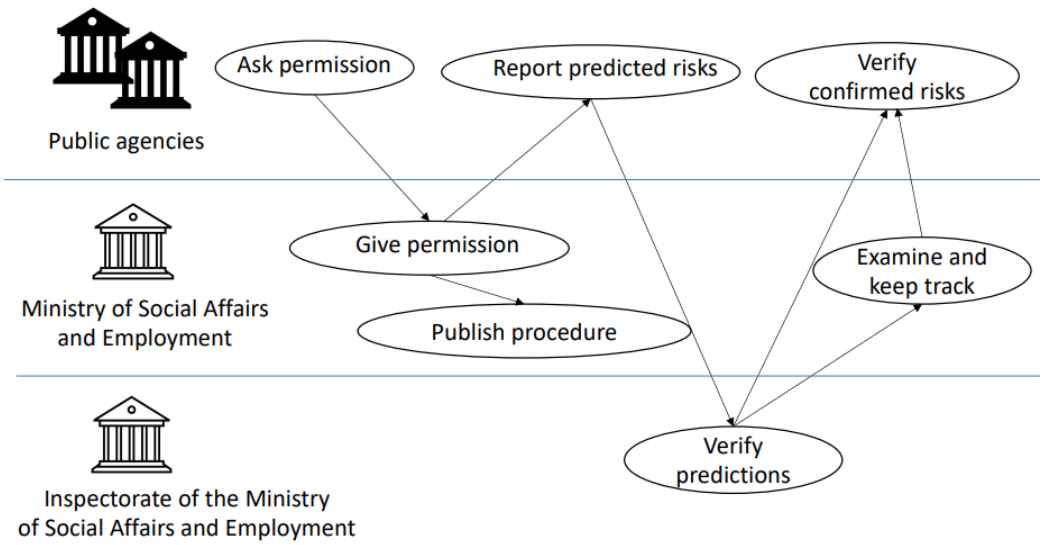
Various municipalities in the Netherlands have been using the SyRi **system to detect welfare fraud more effectively**. SyRi has been developed by the Dutch government and uses various risk indicators from existing governmental systems such as taxes, health insurance, residence, education and many more, in order to detect which addresses hold a higher risk of fraud or misuse of welfare benefits.

Adoption mechanism

- When a governmental agency suspected welfare fraud in a specific neighborhood (benefits, allowances, or taxes), it should cooperate with another agency and ask the Ministry of Social Affairs and Employment to have the SyRI system deployed.

• Municipalities, the Employee Insurance Agency, the social security bank, inspectors of the Ministry of Social Affairs and Employment and the tax authority can ask access to the system

SyRI Process



System working mechanisms and procedures

• SyRI was trained on historical data of residents of Dutch municipalities for patterns of social security fraud

• SyRI produced a prediction on which which citizens of a selected neighbourhood were suspected of welfare fraud

• Positive predictions (suspected welfare fraud) sent to the Inspectorate of the Ministry of Social Affairs and Employment

Verification

• The Inspectorate analyse and verify the predictions, and then report back to the requesting agencies and to Ministry of Social Affairs and Employment

• Ministry of Social Affairs and Employment examines flagged citizens for false positives, and keep confirmed positive risk reports for a maximum of two years.

• The requesting agencies further investigates the frauds, and only if confirmed a sanction could be imposed.

Data used by the system

• Since 2014, SyRi integrated personal data about citizens from several governmental bodies (both central and decentralised):

• work, fines, penalties, taxes, properties, housing, education, retirement, debts, benefits, allowances, subsidies, permits, exemptions, and more

• Pseudonymized data : citizen's names replaced by a unique identifier for each individual, so that it is merge data about citizens from the several data sources used.

• Identifiers are translated back to real names to make sanctions.

Deployment and impact

• SyRI was deployed only in low-income neighborhoods

• Selection bias: more high risks citizens will be found there

• Feedback loop if training data is updated with new fines

• Reinforced stereotyping and negative image of "problem zones"

• → High potential of structural impact

Citizens' perspective

• Citizens were not automatically informed about the investigation

• SyRi's risk reports were inserted in a registry that citizens could view only upon request.

• → If individuals do not know they are investigated, they will not require to check their own data and cannot access the reasons why they have been flagged.

• In practice, when an individual was sanctioned, she could not defend herself

Transparency and reliability

• In 2018, the Ministry of Social Affairs decided that the SyRI risk models should be kept secret otherwise citizens could potentially adjust their behavior accordingly

• Dutch government refused to offer transparency on SyRi's black-box when motions were filled in Parliament (in 2018) and later in Court (in 2020)

• After an independent audit carried by the Netherlands' Organization for Applied Scientific Research research institute, the auditors reported:

  • "The results of the algorithm do not appear to be reproducible"

  • "The risks indicated by the AI algorithm are largely randomly determined," the researchers found.

Detecting social welfare frauds

Feb. 2020

Historical sentence of the Court of The Hague striking down the collection of data and profiling for Social Security fraud (SyRI).

very important

ARTICLE 8

Right to respect for private and family life

1. Everyone has the right to respect for his private and family life, his home and his correspondence.

2. There shall be no interference by a public authority with the exercise of this right except such as is in accordance with the law and is necessary in a democratic society in the interests of national security, public safety or the economic well-being of the country, for the prevention of disorder or crime, for the protection of health or morals, or for the protection of the rights and freedoms of others.

European Convention of Human Rights (ECHR), Art. 8

6.7. The SyRI legislation does not meet the requirement laid down in Article 8, paragraph 2 of the ECHR that interference in the exercise of the right to respect for private life is necessary in a democratic society, that is to say necessary, proportional (proportional) and, in the alternative, in relation to the intended purpose. The court compared the content of the SyRI legislation in the light of the purposes that this legislation serves against the breach of private life that the SyRI legislation makes. It is of the opinion that the legislation does not comply with the 'fair balance' that must exist under the ECHR between the social interest that the legislation serves and the violation of the private life that the legislation produces in order to be able to speak. about a sufficiently justified breach of private life. In doing so, the court takes into account the fundamental principles on which data protection under Union law (the Charter and AVG) is based, in particular the principles of transparency, the purpose limitation principle and the principle of data minimization. She believes that the legislation regarding the use of SyRI is insufficiently clear and verifiable. It is for that reason that the court will declare Article 65 of the SUWI Act and Chapter 5a of the SUWI Decree to be incompatible with this judgment on grounds of conflict with Article 8, paragraph 2 of the ECHR.

6.23. Over time, the ECtHR has classified various interests under the notion of private life and has thereby brought it under the protection of Article 8 of the ECHR. The right to respect for private life also protects a right to personal autonomy, to personal development and self-development and the right to enter into relationships with others and the outside world. According to the ECtHR, the principles of human dignity and human freedom belong to ' *the very essence of the Convention* '. [13] Together with the notion of personal autonomy, they play an important role in determining the scope of the right to respect for private life.

6.24. The right to personal identity and the right to personal development are also mentioned by the ECtHR as sub-aspects of the right to respect for private life. The right to personal identity is also closely related to the right to protection of personal data. Finally, in the case of data processing, the right to respect for private life also affects the right to equal treatment in equal cases and the right to protection against discrimination, stereotyping and stigmatization.

An example

- "Discrepancies" found in data could lead to a risk indication.
- An example is a low usage of running water:
    - Rational: someone who receives house benefits for single actually lives together with someone else at another address
    - Many other plausible causes are excluded by design (broken water meter, saving water, frequent business travels, etc.).
- Any possible data could be tracked for the purpose of the system

Other motivations for the Court decision

- Involved risk of unintentional discriminatory effects (false positives)
- Infringement of fundamental EU principles of data protection:
    - Transparency
    - Purpose limitation
    - Data minimization
    - Contestation

Wider considerations

• The foundations of SyRI lied in an institutionalised distrust towards citizens by a repressive welfare state

• Unmanaged conflict of values: protection of societal values and interests (fraud prevention) vs individuals' rights.

• Public distrust and outrage

• Alternative goals are possible: e.g., predict citizens who have the right to an allowance and did not apply for it

• People need actionable resources, automation for empowering

The childcare-benefits affair in a nutshell

• Since the early 2000's, the Tax and Customs Administration deployed the Fraud Signalling Facility (FSV in dutch language), a non-public risk- profiling system

• 'Blacklist' of 180.000 citizens suspected of fraud

• Many individuals on the blacklist were wrongly identified but still severely damaged.

• Like in Siry, the individuals on the blacklist were not informed and could not defend themselves

• The Tax and Customs Administration suffered from both poor data quality and poor data management.

Issues

• Fraud signals were registered without attributing a weight to them

• Data were kept for too long and too many officials could access the system

• Profiling of groups of citizens during additional checks

• The nationality of childcare-benefit applicants among automatically flagged applications as high-risk (i.e., potentially incorrect and fraudulent)

• When an individual was flagged as a fraud risk, a manual review was conducted by a civil servant, but without explanation on the risk score

Impact

• Tens of thousands of parents and caregivers falsely accused of fraud by the Dutch tax authorities.

• People from ethnic minorities and low-income families disproportionately impacted.

• No answers and explanations given to parents and caregivers identified by the tax authorities as fraudsters, for years.

«Parents and caregivers who were selected by the system had their benefits suspended and were subjected to hostile investigations, characterized by harsh rules and policies, rigid interpretations of laws, and ruthless benefits recovery policies.»

«This led to devastating financial problems for the families affected, ranging from debt and unemployment to forced evictions because people were unable to pay their rent or make payments on their mortgages. Others were left with mental health issues and stress on their personal relationships, leading to divorces and broken homes»

Summary of issues

• Intervention vs prediction: fraud identification and punishment does not remove its cause

• Target-construct mismatch: supporting officials vs repressive policy

• Distribution shifts: massive training vs deployment in neighborhoods

• Limits to prediction: high error rates

• Disparate performance: low-income and minorities

• Lack of contestability: due to opaque processes

• Goodhart's law: “used” to prevent transparency

Chapter 10- Deontology

Normative ethics – schools of thought

• Consequentialism (or Teleological Ethics) focuses on the outcomes an action, investigating morality in accordance with the action's results.

• Deontology defines rules for judging the morality of an action, regardless its consequences, and considering duties and the rights.

• Virtue Ethics focuses on the character of a person/group of persons, identifying virtues and wisdom to deal with virtues conflicts.

A computing professional should...

1. Ethical principles
2. Professional responsibilities
3. Principles of professional leadership (either formal leadership or informal leadership)

(4. Code compliance)

public goods

*public goods regardless of the place whether it will be a company whether it will be an organization a public organization whether it be it will be third sector whether you will be teaching

*impact of data-driven digital Technologies on our life Translation is too long to be saved

تاثیر فناوری های دیجیتال مبتنی بر داده بر زندگی ما

public good is not in contrast with the profit you have an obligation to contribute to the economic well-being of the companies.

in general, organizations in which you will work and at the same time public good is still the most important concern so there is no conflict about these aspects

به طور کلی، سازمان هایی که در آنها کار خواهید کرد و در عین حال خیر عمومی همچنان مهم ترین دغدغه است، بنابراین هیچ تعارضی در مورد این جنبه ها وجود ندارد.

the public code is always the primary consideration of computing professionals

The Code is not an algorithm for solving ethical problems; rather it serves as a basis for ethical decision-making.

Commented [ZK96]: where we could place the fairness quantitative measures
جایی که بتوانیم معیارهای کمی عادلانه را قرار دهیم

Commented [ZK97]: the ontology tries to define for given profession what are your obligations your moral obligations and the professional obligations

هستی شناسی تلاش می کند تا برای حرفه معین، تعهدات اخلاقی و تعهدات حرفه ای شما را مشخص کند.

Commented [ZK98]: virtue ethics where we could place the fairness qualitative assessment which is based on value-based design

اخلاق فضیلتی که در آن می‌توانیم ارزیابی کیفی منصفانه را که مبتنی بر طراحی مبتنی بر ارزش است، قرار دهیم

Commented [ZK99]: از خواندن یک مقاله مطالب مهم زیر استنباط شده است
https://www.acm.org/code-of-ethics

Commented [ZK100]: خیر عمومی با سودی که شما موظف به کمک به رفاه اقتصادی شرکت ها هستید در تضاد نیست

this code doesn't remove your own responsibility and your own decision your autonomy and making decisions

Questions related to these kinds of issues can best be answered by thoughtful consideration of the fundamental ethical principles, understanding that the public good is the paramount consideration

several aspects in which you have an obligation so you have an obligation to take care of the quality of life of people both individually and collectively through the digital tools that you build. you have aspects regarding fundamental human rights and you have aspects obligations regarding protecting the autonomy of people.

The entire computing profession benefits when the ethical decision-making process is accountable to and transparent to all stakeholders. Open discussions about ethical issues promote this accountability and transparency.

Avoid Harms

Examples of harm include unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment.

مثالهایی که در طول دوره بکار برده شده مصادیقی از صدمات اجتماعی هستند:

there are certain number of legal early

- legal disputation with chat GPT for instance for damage of reputation of people which was falsely accused by The Tool of sexual harassment
- a professor in the USA
- wrong usage of public money this regards a major in Australia

the consequences of data aggregation and emergent properties of systems should be carefully analyzed

پیامدهای تجمیع داده ها و ویژگی های نوظهور سیستم ها باید به دقت تجزیه و تحلیل شوند

35:37

Commented [ZK101]: این کد مسئولیت شما و تصمیم شما استقلال و تصمیم گیری شما را حذف نمی کند

Commented [ZK102]: پرسش‌های مرتبط با این نوع مسائل را می‌توان با در نظر گرفتن اصول اخلاقی بنیادی به بهترین نحو پاسخ داد، با درک این که خیر عمومی مهمترین ملاحظات است

Commented [ZK103]: چندین جنبه که در آن شما تعهد دارید، بنابراین شما موظف هستید که از طریق ابزارهای دیجیتالی که می‌سازید، از کیفیت زندگی افراد چه به صورت فردی و چه جمعی مراقبت کنید. شما جنبه هایی در مورد حقوق اساسی بشر دارید و جنبه های تعهداتی در رابطه با حمایت از خودمختاری مردم دارید

Software is considered Ethical Source if it meets the following five criteria:

1. It is freely distributed with source code and can be used or combined with other software without a royalty or fee.

2. It is developed in public and is welcoming of community contributions.

3. Its community is governed by a code of conduct that is consistently and fairly enforced.

4. Its creators have the right to prohibit its use by individuals or organizations engaged in human rights violations or other behavior that they deem unethical.

5. Its creators have the right to solicit reasonable and voluntary compensation from the communities or institutions that benefit from the software.

Broader initiatives on ethics and governance of AI

According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment

• **Human agency and oversight:** AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the loop, and human-in-command approaches

• **Technical Robustness and safety:** AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that unintentional harm can be minimized and prevented.

• **Privacy and data governance:** besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimized access to data.

• **Transparency:** the data, system and AI business models should be transparent. Traceability mechanisms can help achieve this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system's capabilities and limitations.

• **Diversity, non-discrimination and fairness:** Unfair bias must be avoided, as it could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

• **Societal and environmental well-being:** AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully considered.

• **Accountability:** Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate and accessible redress should be ensured.

Australian ethical AI principles 1/2

P1: Human, social and environmental wellbeing: Throughout their lifecycle, AI systems should benefit individuals, society and the environment.

P2: Human-centred values: Throughout their lifecycle, AI systems should respect human rights, diversity, and the autonomy of individuals.

P3: Fairness: Throughout their lifecycle, AI systems should be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups.

P4: Privacy protection and security: Throughout their lifecycle, AI systems should respect and uphold privacy rights and data protection, and ensure the security of data.

• **P5: Reliability and safety:** Throughout their lifecycle, AI systems should reliably operate in accordance with their intended purpose.

• **P6: Transparency and explainability:** There should be transparency and responsible disclosure to ensure people know when they are being significantly impacted by an AI system, and can find out when an AI system is engaging with them.

• **P7: Contestability:** When an AI system significantly impacts a person, community, group or environment, there should be a timely process to allow people to challenge the use or output of the AI system.

• **P8: Accountability:** Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled.

Montreal declaration MONTRÉAL DECLARATION FOR A RESPONSIBLE DEVELOPMENT OF AI

1) WELL-BEING PRINCIPLE

2) RESPECT FOR AUTONOMY PRINCIPLE

3) PROTECTION OF PRIVACY AND INTIMACY

4) SOLIDARITY PRINCIPLE

5) DEMOCRATIC PARTICIPATION PRINCIPLE

6) EQUITY PRINCIPLE

7) DIVERSITY INCLUSION PRINCIPLE

8) CAUTION PRINCIPLE

9) RESPONSIBILITY PRINCIPLE

10) SUSTAINABLE DEVELOPMENT PRINCIPLE

Neglected relevant issues

Example 1: AI and indecent work

Dignity of employees was one of the principles that were listed in the ACM code of conduct. We have a huge problem in AI with laborers.

Human laborers that are especially in the content moderation field work in very bad conditions not only from the economic point of view but also from the psychological point of view.


BUSINESS + TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

This image was generated by OpenAI's image-generation software, Dall E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible.

imagine watching for eight hours a day videos of violence physical violence on minors and either directly remove them from the platform or instruct the algorithms to remove them. It is training helpful in labeling. This is the reality of how high the big companies are facing a Content moderation and in general using work in the global South to label the massive amount of data that they need.

Commented [ZK104]: کرامت کارکنان یکی از اصولی بود ذکر شده بود. ما در هوش ACM که در آیین نامه رفتاری مصنوعی با کارگران مشکل بزرگی داریم.
کارگران انسانی که به ویژه در حوزه اعتدال محتوا هستند، نه تنها از نظر اقتصادی بلکه از نظر روانی نیز در شرایط بسیار بدی کار می کنند.
تصور کنید که به مدت هشت ساعت در روز ویدیوهایی از خشونت فیزیکی روی کودکان زیر سن قانونی تماشا کنید و یا مستقیماً آنها را از پلتفرم حذف کنید یا به الگوریتم ها دستور دهید آنها را حذف کنند. این آموزش در برچسب زدن مفید است. این واقعیتی است که شرکت‌های بزرگ تا چه حد با تعدیل محتوا مواجه هستند و به طور کلی از کار در جنوب جهانی برای برچسب‌گذاری حجم عظیمی از داده‌های مورد نیاز خود استفاده می‌کنند.

Example 2: copyright

for example Copyright:

there is a massive privatization of the work of thousands of people millions of even millions of people because the new models take training

whatever they can whatever they are able to be whatever data they are they are able to even through commercial contracts so not only through data that is already on the web.

Other relevant issues

• Social responsibility (e.g., work conditions)

that is evaded by the big players in AI.

• Liability

What is the responsibility for a crash of a car that is partially driven by AI? we don't know yet. but normally it seems that normal liability principles do not apply in their field and this is not at the moment right.

• Mass appropriation of resources

• Socialization of risks and damages (e.g., environment)

the huge impact of large models, for instance a study that was made on an Alexa, highlighted that just to train a system one time, the system like Alexa is equivalent to go back and forth from New York to Beijing for more than 500 times in terms of emissions. you can imagine with ChatGPT there is some data, it's quite impressive.

• Proper taxation and economic redistribution

redistribution huge amounts of economic power that is equivalent to the size of a state economy. taxation very little there is what's here in which Amazon paid zero in U.S taxes.

• Innovation and competition

are killed at the moment by the major players even by direct acquisition.

• Concentration of power

which is a consequence of the huge economic wealth that has been accumulated.

• Impact on democratic processes

Commented [ZK105]: به عنوان مثال حق چاپ: خصوصی سازی گسترده ای از کار هزاران نفر و حتی میلیون ها نفر وجود دارد زیرا مدل های جدید آموزش می بینند. هر چه که می توانند هر چه می توانند باشند، هر چه می توانند باشند، حتی از طریق قراردادهای تجاری، نه تنها از طریق داده هایی که از قبل در وب هستند.

Commented [ZK106]: توسط بازیکنان بزرگ در هوش مصنوعی طفره می رود.

Commented [ZK107]: مسئولیت تصادف خودرویی که تا حدی توسط هوش مصنوعی هدایت می شود چیست؟ ما هنوز نمی دانیم اما به طور معمول به نظر می رسد که اصول تعهدات عادی در زمینه خود اعمال نمی شود و در حال حاضر درست نیست

Commented [ZK108]: اجتماعی شدن محیط خطر و آسیب اول از همه به دلیل تأثیر بسیار زیاد مدل‌های بزرگ، به عنوان مثال، مطالعه‌ای که بر روی یک الکسا انجام شد، نشان داد که فقط برای آموزش یک سیستم یک بار، سیستمی مانند الکسا معادل رفت و برگشت از نیویورک به پکن برای اطلاعات بیشتر است. از نظر انتشار بیش از 500 برابر می توانید داده هایی وجود دارد، بسیار ChatGPT تصور کنید که با چشمگیر است.

Commented [ZK109]: مالیات مناسب و بازتوزیع اقتصادی: بازتوزیع حجم عظیمی از قدرت اقتصادی که معادل اندازه یک اقتصاد دولتی است. مالیات بسیار کمی وجود دارد که در اینجا آمازون مالیات های ایالات متحده را صفر پرداخت کرده است.

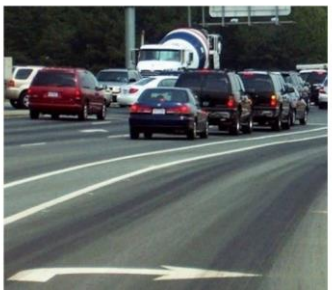
Commented [ZK110]: رقابت نوآوری: در حال حاضر توسط بازیگران اصلی حتی با خرید مستقیم کشته می شوند

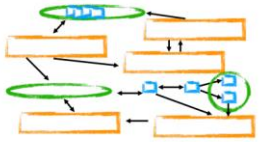
Commented [ZK111]: تمرکز قدرت: که نتیجه ثروت عظیم اقتصادی است که انباشته شده است

Commented [ZK112]: تأثیر بر فرآیندهای دموکراتیک ما: با ظهور هوش مصنوعی مولد چه اتفاقی برای اینفوسفر ما خواهد افتاد؟ ما واقعاً نمی دانیم، 30 سال از تاریخ وب طول کشید تا وب را بسازیم که می شناسیم و این در حال حاضر با ابزارهایی که می توانند به روشی بسیار مقیاس پذیر می توانند هزاران سند جعلی را برای یک هدف معین بنویسند مشکل ساز است. نمی دانم چه اتفاقی برای فرآیندهای دموکراتیک ما خواهد افتاد.

What will happen to our infosphere with Advent of generative AI? we really don't know, it took 30 years of history of the web to build the web that we know and this is already problematic with tools that can be in a very scalable way can write thousands of fake documents for a given purpose we really don't know what will happen to our Democratic processes.

Complex socio-technical systems





From the technical point of view of an engineer, they should design and develop a highway and should not look only at the capacity of the highway's number of lanes or the type of concrete. should also look at the societal impact and those disciplines. This aspect is very mature. So there are impact studies and risk studies that are given before you build such an infrastructure and also to avoid problems like this.

Commented [ZK113]: از نظر فنی یک مهندس باید بزرگراه را طراحی و توسعه دهند و فقط به ظرفیت تعداد خطوط بزرگراه یا نوع بتن نگاه نکنند. همچنین باید به تأثیرات اجتماعی و آن رشته ها نگاه کرد. این جنبه بسیار پخته است. بنابراین مطالعات تاثیر و مطالعات ریسک وجود دارد که قبل از ساخت چنین زیرساختی و همچنین برای جلوگیری از مشکلاتی از این دست انجام می شود

Data Ethics notes

Where privacy is fun

Big data raises several important concerns:

- Social acceptability and adoption
- Ethical implications
- Responsibility and accountability of data scientists

"Verificare il libero che ha scritto il professore" requires clarification. Past exams may not provide a clear roadmap due to the diversity of questions. Achieving the minimal threshold in each question is necessary for passing. A legal system is influenced by its geographic location and temporal context. Despite these changes, certain similarities may be identified across countries, often stemming from historical connections.

The Role of International Laws in Data Protection

The General Data Protection Regulation (GDPR), stemming from the EU, must be ratified by national parliaments. International law and treaties form the bedrock of such legal provisions at a global level. Notably, before the implementation of the first data directive regulations, the EU primarily relied on international law for data protection.

مقررات عمومی حفاظت از داده ها (GDPR) که از اتحادیه اروپا سرچشمه می گیرد، باید توسط پارلمان های ملی تصویب شود. حقوق بین الملل و معاهدات بستر چنین مقررات حقوقی را در سطح جهانی تشکیل می دهند. قابل ذکر است، قبل از اجرای اولین مقررات دستورالعمل داده، اتحادیه اروپا در درجه اول برای حفاظت از داده ها به قوانین بین المللی متکی بود.

The Impact of Supranational Legislation

When a supranational entity such as the EU legislates a law, member states are required to adjust their sovereignty to accommodate these new rules. This necessity arises from past agreements and the imperative to regulate regional issues that extend beyond national confines. The EU data directive, for instance, was enacted when all EU nations began extensive data sharing. Similarly, cybercrime, which cannot be contained within national borders, mandates international agreements for its effective resolution.

تأثیر قانونگذاری فراملی

هنگامی که یک نهاد فراملی مانند اتحادیه اروپا قانونی را وضع می کند، کشورهای عضو ملزم به تنظیم حاکمیت خود برای تطبیق با این قوانین جدید هستند. این ضرورت ناشی از توافقات گذشته و ضرورت تنظیم مسائل منطقه ای است که فراتر از مرزهای ملی است. به عنوان مثال، دستورالعمل داده های اتحادیه اروپا زمانی تصویب شد که همه کشورهای اتحادیه اروپا به اشتراک گذاری گسترده داده ها را آغاز کردند. به همین ترتیب، جرایم سایبری که نمی تواند در داخل مرزهای ملی مهار شود، توافقات بین المللی را برای حل موثر آن الزام می کند.

From Inter-State Agreements to Direct Regulations

All international agreements demand tangible implementation of laws, as these inter-state agreements do not directly apply to citizens. The EU, initially designed for fostering agreements between states, has evolved due to the emergence of regulations that directly apply to citizens, effectively bypassing national parliaments.

- Directive: This is addressed to states, to be enacted by national parliaments, offering a degree of flexibility. It paves the way for initial harmonization.
- Regulation: This applies directly and immediately to all citizens of member states.

The Goal of EU Market Harmonization

The EU aspires to create a market that facilitates a seamless flow جریان یکپارچه of goods and people. However, the common market can be hampered by borders, taxation, and most

significantly, regulations. For instance, a car produced in Germany can't be sold in Spain if it doesn't conform to Spanish regulations. Thus, a key objective of the EU is to harmonize(coordinate هماهنگ کردن) the laws of the common market among its member states.

However, the EU's harmonization primarily relates to the market and doesn't cover societal issues and social structures unless they impact the market. While the EU safeguards (protect) fundamental rights, it legislates these rights primarily because they harmonize the market. The current proposals on AI, for instance, focus solely on high-risk applications to facilitate market access while also ensuring the protection of citizens' rights.

Variations in Pre-EU Data Protection Laws

Before the introduction of EU data protection law, EU member countries had disparate regulations, or in some cases, none at all.

During negotiations, EU diplomats voiced concerns about their countries' vulnerabilities and advocated for their existing laws. This is largely because any systemic change carries costs and transition periods. Countries with established regulations have an advantage over those needing to update theirs, making EU regulation harmonization is a challenging task due to the necessity of widespread agreement.

تغییرات در قوانین حفاظت از داده های قبل از اتحادیه اروپا

قبل از معرفی قانون حفاظت از داده‌های اتحادیه اروپا، کشورهای عضو اتحادیه اروپا مقررات متفاوتی داشتند یا در برخی موارد اصلاً مقرراتی نداشتند.

در جریان مذاکرات، دیپلمات های اتحادیه اروپا در مورد آسیب پذیری کشورهای خود ابراز نگرانی کردند و از قوانین موجود آنها حمایت کردند. این عمدتا به این دلیل است که هر تغییر سیستمی هزینه و دوره های انتقالی را به همراه دارد. کشورهایی که دارای مقررات تثبیت شده هستند نسبت به کشورهایی که نیاز به به روز رسانی مقررات خود دارند، برتری دارند، به دلیل لزوم توافق گسترده، هماهنگ سازی مقررات اتحادیه اروپا یک کار چالش برانگیز است.

Direct Implementation and Preventive Regulations in the EU

اجرای مستقیم و مقررات پیشگیرانه در اتحادیه اروپا

Certain areas in the EU, like technical standards, see regulations directly implemented. One recent development is the Artificial Intelligence Act (ACT), a pre-emptive regulation drafted to prevent divergence before countries start individually regulating AI.

برخی از مناطق در اتحادیه اروپا، مانند استانداردهای فنی، مقرراتی را مشاهده می کنند که مستقیماً اجرا می شوند. یکی از پیشرفت‌های اخیر، قانون هوش مصنوعی (ACT) است، یک مقررات پیشگیرانه که برای جلوگیری از واگرایی قبل از اینکه ایالت‌ها به طور جداگانه تنظیم کنند، پیش‌نویس شده است.

ACTs are special powers granted to the Commission based on treaties to enact laws on as-yet unregulated areas. For an ACT to be justified, the Commission must demonstrate the necessity of regulation before the member states implement their own.

ACT ها اختیارات ویژه ای هستند که توسط معاهدات به کمیسیون برای قانون گذاری در زمینه هایی که هنوز تنظیم نشده اند اعطا می شود. برای اینکه یک ACT توجیه شود، کمیسیون باید قبل از اجرای قوانین خود توسط کشورهای عضو، ضرورت تنظیم مقررات را نشان دهد.

This shift in regulatory approach significantly alters the power dynamic between the Commission and the member states. Brussels, rather than Paris or Berlin, has become the primary legislative center. The ACT, similar to regulation, is immediately legally binding, thereby pre-empting pan-European discussions.

این تغییر در رویکرد نظارتی به طور قابل توجهی پویایی قدرت بین کمیسیون و کشورهای عضو را تغییر می دهد. بروکسل، به جای پاریس یا برلین، به مرکز اصلی قانونگذاری تبدیل شده است. ACT، مشابه مقررات، فوراً از نظر قانونی الزام آور است و در نتیجه مانع از بحث های پاناروپایی می شود.

Yet, national laws can sometimes conflict with EU regulations. The EU Court of Justice oversees how states implement EU laws. In accordance with EU treaties, states must ensure their laws align with those of the EU.

با این حال، قوانین ملی گاهی اوقات می تواند با مقررات اتحادیه اروپا در تضاد باشد. دیوان دادگستری اتحادیه اروپا بر نحوه اجرای قوانین اتحادیه اروپا توسط کشورها نظارت می کند. مطابق با معاهدات اتحادیه اروپا، کشورها باید اطمینان حاصل کنند که قوانین آنها با قوانین اتحادیه اروپا هماهنگ است.

While the EU lacks an emergency rule mechanism, sub-EU institutions can implement such rules. Data protection, particularly concerning large platforms, requires stringent (precise, exact دقیق) control due to its widespread impact.

در حالی که اتحادیه اروپا فاقد مکانیزم قوانین اضطراری است، نهادهای زیر اتحادیه اروپا می توانند چنین قوانینی را اجرا کنند. حفاظت از داده ها، به ویژه در مورد پلتفرم های بزرگ، به دلیل تأثیر گسترده آن به کنترل دقیق نیاز دارد.

Co-Regulation: A Balance of Hard and Soft Laws

تنظیم مشترک: تعادل قوانین سخت و نرم

Not all rules originate from countries or national bodies. Some are created from the bottom up, forming 'soft laws'. These are agreements or spontaneous convergences between parties, different from 'hard laws' in their system impact.

همه قوانین از کشورها یا نهادهای ملی سرچشمه نمی گیرند. برخی از آنها از پایین به بالا ایجاد می شوند و "قوانین نرم" را تشکیل می دهند. اینها توافق‌ها یا همگرایی‌های خود به خودی بین طرفین است که از نظر تأثیر سیستمی با «قوانین سخت» متفاوت است.

Hard laws are compulsory, necessitate a lengthy legislative process, and are prone to political influence. They can also be rigid(hard , difficult), making amendments time-consuming. Often drafted by parliamentarians who may lack topic expertise, these laws can occasionally be sub-optimal.

قوانین سخت اجباری هستند، نیازمند یک فرآیند طولانی قانونگذاری هستند و مستعد نفوذ سیاسی هستند. آنها همچنین می توانند سفت و سخت باشند، که باعث می شود اصلاحات زمان بر باشد. این قوانین که اغلب توسط نمایندگان مجلس که ممکن است فاقد تخصص موضوعی باشند تهیه می‌شوند، گاهی اوقات می‌توانند از حد مطلوبی برخوردار نباشند.

Soft laws, on the other hand, are crafted by stakeholders who understand the problem and the required solutions. Despite being faster to enforce and catering to a broader or more specific sector, soft laws are weaker as they lack the binding power of state law.

از سوی دیگر، قوانین نرم توسط ذینفعانی تدوین می شوند که مشکل و راه حل های مورد نیاز را درک می کنند. قوانین نرم علیرغم اجرای سریعتر و ارائه خدمات به بخش وسیع تر یا خاص تر، ضعیف تر هستند زیرا فاقد قدرت الزام آور قانون دولتی هستند.

The EU attempts to blend these two systems through a 'co-regulation' strategy. In this scenario, breaching a code of conduct could constitute(organize) a violation of the law, marrying the flexibility of soft law with the power of hard law.

اتحادیه اروپا تلاش می‌کند تا این دو سیستم را از طریق یک استراتژی «نظارت مشترک» ترکیب کند. در این سناریو، نقض قوانین رفتاری می‌تواند نقض قانون باشد و انعطاف‌پذیری قانون نرم را با قدرت قانون سخت تطبیق دهد.

According to the lex specialis principle, a new law can coexist with existing ones. This principle states that a general law can operate alongside a more specific one, with the latter prevailing(dominant) when the two conflict.

طبق اصل lex specialis، یک قانون جدید می تواند با قوانین موجود همزیستی داشته باشد. این اصل بیان می‌کند که یک قانون کلی می‌تواند در کنار قانون خاص‌تر عمل کند، و قانون دوم در هنگام تضاد این دو غالب باشد.

Rule Making

The Case of Smoking

Consider a scenario where a state aims to reduce smoking among its population. Several strategies might be used:

1-Increasing Product Costs: By **raising taxes** on tobacco products, the state uses market forces to shape societal behavior.

2-Implementing Non-Smoking Area Sanctions: Here, the **law** is used as a **tool**. However, **enforcing such laws can be challenging**, and they may clash with individual rights to self-determination. For instance, a person may still choose to smoke in private settings like their home.

3- Changing Attitudes: If people are **personally convinced** that they should not smoke, they are less likely to do so.

Law and Technology

Do New Technologies Necessitate New Laws?

The question of why we choose to regulate technology and whether it's necessary arises frequently. While the advent of new technology often sparks the need for fresh legislation, creating a new law every time could lead to conflicts.

For instance, defamation(accusation تهمت زدن) in person and online defamation(هتک حرمت در فضای آنلاین) could be handled by the same law. When the EU decided **to regulate e-commerce**, it limited its legislation, allowing most pre-existing laws to apply. The rationale was to mirror real-world laws - for instance, online purchases should have replacement policies similar to physical stores.

Liability Shift from Publishers to Internet Service Providers

تغییر مسئولیت از ناشران به ارائه دهندگان خدمات اینترنتی

Two landmark cases highlighted this shift in liability. In the **case of Cubby INC**, the court ruled that a third-party daily newsletter was similar to a newspaper, hence the distributor couldn't be held liable for distributing defamatory content. In Stratton vs. Oakmont, the court ruled that if content was moderated, the platform was acting like a publisher, thus making them liable for defamation. This lead to platforms choosing not to moderate to avoid liability, an unfavorable outcome. Subsequent laws addressed this issue, holding platforms liable if they were aware of or promoted defamation.

Choosing the Right Regulatory Approach:

Prescriptive vs Preventive Paradigm In the digital context, law enforcement often proves more challenging than in the physical world. Crimes may go undetected for a long time, and their international scope complicates matters further. Solely relying on a prescriptive approach may not be effective. A preventive paradigm, which emphasizes behaviors and technology to stop infringements before they occur, might be more successful. The 'by design' approach incorporates these preventive measures directly into devices or technology. For instance, the General Data Protection Regulation (GDPR) has a dedicated section for privacy by design. A balanced mix of enforcement methods, including tech-embedded rules and default behavioral laws, is crucial.