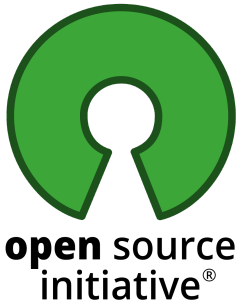
DATA ETHICS AND PROTECTION

A.Y. 2021/2022

Politecnico di Torino

MSc Data Science and Engineering

These notes cover the whole program.



Disclaimer

The content is the result of the writer's interpretation, therefore there are no guarantees about the correctness and completeness.

Be aware that the content of the course changes each academic year, therefore most part of this document will become out of date at the end of A.Y. 2021/2022, unless it will be updated by other students.

TABLE OF CONTENTS

Other contributors

- **Leonardo Tredese**
- **Giuseppe Concialdi (grazie per lo spritz)**
- **Christian Montecchiani**
- **Edgar Gaytán**

Data protection - GDPR

Introduction - Main pillars

To protect personal data, the European Union issued in 2016 the GDPR (General Data Protection Regulation). It is not a **directive** (i.e. each State can implement it in a different way) but it is a **regulation** (a single State cannot modify it).
The GDPR regulates only personal data, other kinds of data are not taken into account.
Its main pillars are:

- ***data-centric model***: GDPR considers data as the core of the online world.
- ***procedural approach*****:** it is something that could be chosen each time (different procedure depending by the context). It is the opposite of a strictly legal approach (same approach for every situation).
- ***tech neutral approach***: the typical example for this point is the two-factors authentication (2FA). GDPR does not explain which technologies must be used to implement 2FA since they change faster than the law. Instead, whenever you make a decision about implementation of 2FA, you need to be able to justify it with respect to GDPR.
- ***risk management***: each data controller must produce a risk assessment document, where all the risks of managing certain data are presented.
- ***legal basis****:* the consent is not the only way to process and analyze personal data, there are other legal basis. For example:
    - **Law**: to start a criminal procedure against a person you do not need his/her consent to use his/her personal data.
    - **Medical Assistance**: a doctor does not need the consent when he is trying to save a life.

The professor reported the *"Google Spain case"*: a man was convicted for a crime, but after his punishment was completed, the search engine continued in presenting articles about his crime. So the right to be forgotten goes against the freedom of speech, and each case is different.

Article 1 - Subject-matter and objectives

The GDPR regards physical (natural) persons only, thus it is not applicable to legal entities (such as companies). It concerns the processing of personal data. Its main aim is to protect EU citizens from any decision taken by tech companies which can influence their lives. Article 1 also protects the free movement of data within the European union.

Article 3 - Territorial scope

It is not important if the personal data processing takes place in the EU or not, each controller must have an establishment in the Union.
This point goes against the *territorial principle:* if you visit a website based in Kazakhstan, technically you are under Kazakh jurisdiction. Therefore GDPR should not be applicable. Instead, the EU forces a company to have an establishment in the Union, so that the controller of personal data processing can be controlled and sanctioned. The example brought up during the lecture was the cookies acceptance: in Kazakhstan they do not need to ask you for your consent to use cookies, but if the website is visited from the EU they need to ask.

Article 4 - Definitions

Personal data definition: any information relating to an identified or identifiable (directly or indirectly) natural person.
Data processing definition: it is any operation performed on personal data such as collection, consultation, destruction, transmission, etc…

An example is the IP address: for Italian law, you need a law enforcement entity to ask the name associated with a given IP address, while in Germany even law enforcement is not enough to access this information. IP addresses are considered personal data by GDPR (either static or dynamic), while from the professor's point of view only the static IP is personal data.

Problem with GDPR: it is too restrictive, it can sometimes stop some useful research activities. The solution to this problem is data anonymization, but this process is not well defined and, besides, each case must be analysed on its own. Note that anonymous data are not under the GDPR scope.
NB: aggregated data are not personal data.

Recital: it is the “background” of an article, it explains why that article has been inserted. For example, the recital of article 26 states that data which could identify a person in the future must be protected.

Personal data main principles

- *Lawfulness, Fairness and Transparency*: personal data should be processed lawfully (i.e. according to law), fairly and in a transparent manner in relation to the **data subject**.
- *Purpose limitation*: personal data should be collected for specified, explicit and legitimate purposes. For example, by the Italian law, after 10 years that your purpose is concluded, you must destroy any kind of information about a person involved. So this tells us that **data retention** is really important, since it is not easy to delete every day data from ten years ago, if we have a huge amount of data.
- *Data minimisation*: personal data should be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.
- *Accuracy*: personal data should be accurate and kept up to date.
- *Storage limitation:* personal data should be kept in a form which permits identification of data subjects for no longer than it is necessary for the purposes for which the personal data are processed.
- *Integrity and confidentiality*: personal data should be processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage.

From these principles, we can infer that a controller needs to define a **data retention process**. Actually, defining the life cycle of data management is quite difficult. Every single day we need to delete some data and their backup. Therefore, data minimisation is a business, less data less risk.

Accountability principle

The controller shall be responsible for and be able to demonstrate compliance with the above principles. The accountability principle requires you to take responsibility for what you do with personal data and how you comply with the other principles.

You must have appropriate measures and records in order to be able to demonstrate your compliance. Before GDPR there were strict rules, now if you take a decision against GDPR but your thesis is convincing then it can be approved. Actually, this is a very rare case, but it is a singularity with respect to rules present before.

GDPR pillars - more practical

- *Records of processing activities*: each controller shall maintain a record of processing activities under its responsibility (which data I collect, what I do with data, for how long I store data, etc…)
- *Data processors*: where processing is to be carried out on behalf of a controller, the controller shall use only processors providing sufficient guarantees to implement appropriate technical and

organisational measures (for example PoliTo uses Zoom and BigBlueButton for video conferences, those are external entities out of the controller's control). Data processors and data controller need to make an **agreement** (data processing agreement).
- *DPIA*: it is a document to carry out the risk assessment task. In this document, all the risks for the data subject resulting from the personal data processing are reported. It presents an assessment of the impact of the processing operations on the protection of personal data.
- *Data breach*: in the case of personal data breach, the controller shall notify it to the competent supervisory authority, not later than 72 hours.

Different kinds of personal data

GDPR specifies different kinds of personal data based on the legal basis you need to process them.
A legal basis can be, but is not limited to, a law or the legitimate interest of a company. The **legitimate interest** is a balance between user and company interests.

Criminal records data

This is a special category of personal data because you can process them only if there is a law that allows you to do that. In Europe, after GDPR, companies cannot ask for criminal records of their employees. In fact, also if the employee provides to the company the criminal records, they are not allowed to process that data.
In the USA, instead, criminal records are publicly available.

Health data

This is another special category. In this case, a doctor does not need the explicit consent to process this kind of data if the purpose is saving the life of the patient.

Article 7 - Conditions for consent

- The controller must be capable of demonstrating that the user gave the consent of processing his/her data (e.g. with a log file).
- The policy must be presented in an accessible manner (i.e. it should be easy to be read by a normal user). Actually, this is an issue since nowadays they are quite difficult to be read and understood. The majority of countries (such as the USA, China, the EU, …) agree that the controller must be clear to the user in explaining how he/she will use the subject's data.
- Child's consent: the minimum age to express this consent is from 13 up to 16 years old (it depends on single states). In Italy this age is 14.
-

GDPR - BALANCING TESTING

The interest of the data controller should be balanced with the impact on the data subjects. To carry out a correct balancing test three main points need to be followed:

- Assessing the **legitimate interest** of the controller (lawful, sufficiently clearly articulated, real and present)

- **Impact** on the **data subject** (nature of the data, methods of data processing, reasonable expectations of the data subject, the status of data subject/controller).

- Additional safeguards to prevent any undue impact on the data subjects.

GDPR Individual's rights

- *Right to access*: information about personal data processing, the purpose, what kind of data, the period of storage.
- *Right to rectification*: correction of inaccurate personal data concerning him/her, without delay.
- *Right to erasure*: right to be forgotten, to erase all personal data if not necessary anymore or if the user withdraws consent.
- *Right to restriction of processing*: if data accuracy is contested, unlawful or not needed anymore.
- *Right of data portability*: to receive user's concerning personal data, in a structured format. Example: moving chats from WhatsApp to Telegram. The problem, nowadays, is that platforms do not provide a useful format to export your data, for example you cannot move your social account from a platform to another one easily.
- *Right to object*: stop processing personal data on request, unless the controller demonstrates compelling reasons overriding the individual's interests and rights.

Any kind of violation of GDPR is fined with a sum up to 20 millions euros or up to 4% of the global turnover of a company.

Immuni study case

Immuni app was the contact tracing system developed by the Ministry of Health in Italy for COVID-19 pandemic. The issue was to find a balance between privacy and tracing a person. Despite the purpose of the system was to preserve the public health, and users installed this app on voluntary purpose (so there is the users' consent), the government decided not to use geolocation in order to not violate GDPR. Therefore, the contact tracing works only with bluetooth connection, and actually it didn't work appropriately. Another weird event is that, to approve this system, they issued a legislative decree, also if they had users' consent and health issues as legal basis to process that kind of data.
That's primarily because data from Immuni are **pseudonymised** data, not **anonymised**.
The first category is composed of that kind of data which can be attributed to a natural person through the use of additional information, thus it falls within the scope of the GDPR. The second category, instead, are information that does not relate to an identified or identifiable natural person and prevent/disallow identification of the data subject also with additional data. They do not fall within the scope of the GDPR. Immuni is a clear example of how privacy can limit technology.

Limit risk for stakeholder

In order to limit risks, they used:

- consent of the data subject
- temporary of the measure
- retention period based on necessity (all data will be deleted at the end of the pandemic period)
- location data is neither necessary or recommended

This is a real scenario in which the GDPR was respected.

GDPR compliance

In order to be GDPR compliant, when you process personal data, you should:

- analyse what you collect and where data is stored with the *processing activity register*. In this register you should insert all relevant information for each processing activity. You need to be very clear about what you do with data.
- check if the time you store personal data is relevant. If not, remove data (risk minimization).
- Inform data subjects about how they can modify or delete their data (privacy policy web page).
- Monitor who has access to personal data.

GDPR task distribution

The stakeholders of GDPR are the following.

- *Data subject*: a **natural person**, resident of European Union countries, the subject of data.
- *Data controller*: a **legal entity** (e.g. a company) owner of the personal data processing.
- *Data protection officer*: person appointed by the data controller responsible for overseeing data protection practices.
- *Data processor*: subject that processes data on behalf of the controller (e.g. Google).
- *Data authority*: public institution monitoring implementation of the regulations in the specific EU member country (one for each state).

NB: GDPR can fine only legal entities, people can break GDPR rules only in a criminal way.
Data processor and controller need to make a *data processing agreement* to define responsibilities. One of the risk mitigation principles is to choose a data processor with appropriate security standards and specify that in your risk assessment document.

Sanctions are only towards the data controller. But if the fault was of the processor, the controller can ask the processor for damages.

Data processor

The decision-making power over data processing is the criterion for distinguishing between controller/processor. The processor must provide sufficient guarantees to implement appropriate technical and organisational measures in such a manner that processing will meet the requirements of GDPR (Article 28).
The data processing agreement should specify:

- the subject-matter and duration of the processing
- the nature and purpose of the processing
- the type of personal data and categories of data subjects
- the obligations and rights of the controller

Sub-processor

The processor can engage another processor (**sub-processor**) with a prior specific or general written authorisation of the controller. The initial processor remains fully liable to the controller for the performance of the sub-processor's obligations.

Joint-controller

Two or more controllers jointly determine the purposes and means of processing. There exists a formal agreement between the joint-controllers, which determines respective responsibilities and obligations.

Data protection officer

This entity is the point of contact between controller and authority. It should be independent of the data controller and cannot write the risk assessment, but he/she only gives his opinion. If the data protection officer depends on the controller, the risk is that the GDPR is stretched to fit a bad personal data processing.

By the accountability principle, this entity is mandatory for:

- Public authorities or bodies.
- Data controller/processor whose core activities consist of processing operations which require regular and systematic monitoring of data subjects on a large scale or special categories of personal data (e.g. criminal records).

He/she must be an expert in national and European data protection laws and have an in-depth understanding of the GDPR. He/she must also understand the processing operations carried out and have knowledge about information technologies and data security. Furthermore, he/she should have knowledge about the business sector and the organisation, and have the ability to promote a data protection culture within the organisation (thus he should be independent).

Article 32 - Security measures

This article of the GDPR expresses all the security obligations of the data controller/processor. It does not report an exhaustive list of minimum security requirements, instead, it is strictly related to the accountability principle. In fact, the data controller needs to justify every security decision taken.
In Italian law, before the GDPR, there was a list of minimum requirements that a system that processes personal data should have had. This is unfeasible since the technologies change too fast with respect to the laws.
ENISA (European Union Agency for Cybersecurity) helps to define standards and keep them up to date, used to assess whether a system is GDPR compliant or not.

ENISA report

ENISA identifies the *technical* measures:

1. Cybersecurity culture: most data breaches happen because of human errors, so you need to teach all the employees how not to be a victim of phishing attacks.
2. Training courses: they should be *customized* for each company, and focused on real scenarios.
3. Relations with third parties: a data controller must choose data processors carefully since the data controller will be considered responsible for the entire process.
4. Data breach procedure: a company must prepare a plan to stick to in case of a data breach since the time to react is very short (72 hours).
5. Security access to IT systems: use passphrases (little-used words in sequence) and avoid reusing passwords.
6. Device safety: encrypt all devices given to employees, in case of loss it is a data breach if they are not encrypted.
7. Corporate network security: implementation of firewalls and constant review of all the solutions that allow remote access to the corporate network.
8. Physical corporate security: be careful about physical access to the company's devices.
9. Backup security: it has to be done on a regular basis, automatically and immediately usable.
10. Cloud: check how the cloud is backed up, authentication tools, etc.
11. Websites security: carry out security tests on a regular basis.
12. Search and share information: in a system, it is the key that must be private, not the implementation details.

ENISA identifies these *organizational* measures:

1. Organizational model
2. Privacy policies
3. Policies
4. Privacy by design
5. Consents
6. Documentation

The important ones will be related later on.

Takeaway on security

Here is reported the principles of security expressed in the GDPR:

- *Standards*: good procedures, schemes, and toolkits, and measures to prevent and respond to incidents.
- *Report*: raise awareness of specific risks, periodically.
- *Cooperation*: between providers and authorities, between several and international authorities.
- *Timing*: get things done as soon as possible, in order to avoid a possible lack of awareness of cybersecurity threats.

Real case example: Rousseau platform of political party “Movimento 5 Stelle”
This platform is used by the party to make polls among its participants. It made it possible to know the vote of each person, and had a lot of security issues. Then, the party was fined with respect to article 32 for “inadequate security measures”. The amount of the fine was 50.000€, which is low because the organisation is not lucrative.

Risk assessment

Three steps strategy: preliminary analysis, data strategy and data management, data minimization and by-design approach.

Google Street View case

These concepts are explained by a real example, the Google Street View case (2006).
Google is composed of three main departments: engineering, marketing and lawyering. The main decision power was given to engineers. This, sometimes, led to failures (for example in Google Glass projects marketing said that it was a bad idea, but engineers went on anyway).
In the Street View project, they aim to take a picture of the whole world using a car with cameras on the top. They also keep note of all WiFi’s SSIDs and associate them with geographical coordinates, so that they can track an Android phone even if the geolocation is deactivated. Furthermore, since in 2006 there were a lot of open WiFi, they connected to those networks and sniffed the traffic for a restricted period of time. Thus, they had a lot of legal problems.
Moral of the story: the risk assessment is really important, and it has to be done before the very start of the project, involving all the departments of the company.

3-stages model

- General assessment with a circular approach, i.e. repeating the same assessment after a defined period of time.
- Formal assessment
- Prior consultation (most important): before starting any kind of project which involves personal data processing, you should ask the authority if the data processing strategies are GDPR compliant. When you have a high risk assessment the authority can reply in three ways:
    - letting you start without any restriction (if the risk is not really high or you have adopted appropriate mitigation measures);
    - letting you start after you have filled security lacks;
    - stopping your processing because the risk cannot be mitigated.

DPIA

It is a document, the data protection impact assessment. There are many standard templates, you can use whatever you want but keep in mind the accountability principle. During the lesson, the French model was presented. It is composed of 4 main parts: context, fundamental principles (e.g. what is the legal basis making the processing lawful), risks, and validation.
Key point: you must describe the impact of the data processing on the data subject, and obtain a prior opinion of the data subject.

If you underestimate the risk, the data protection authority evaluates if you did it intentionally and in a fraudulent way, so this creates uncertainty in the hand of the consultant.
In GDPR, data protection authorities act like a *common law system*, which is in contrast to our system (*civilian law* system). In a civilian law system, judges can only interpret laws, while in a common law system, if a law does not exist, a judge can create it in order to punish a criminal.

Data protection by design and by default

You should think about personal data protection from the very beginning of your project so that the costs needed to achieve security will be reduced. Companies should design a new way to inform the user about personal data processing, but nowadays they try to hide how they process data. If you change the way of communication, maybe users could express real consent.
An example of what happens today are *dark patterns*: it is really difficult to find the button needed to close your Amazon account, and this button does not let you close your account, instead it is used to contact customer support. Then you have to ask them to close your account.

Data breach

Article 4 of GDPR contains also the definition of a data breach. It is a security breach that accidentally or unlawfully results in the destruction, loss, alteration, unauthorized disclosure of, or access to, personal data that has been transmitted, stored, or otherwise processed.
NB: if no personal data are involved then it is not a data breach, it is only an IT accident.
GDPR finds three types of data breach:

- confidentiality breach (i.e. unauthorised access): also losing a not encrypted device of the company, if it contains personal data;
- availability breach (i.e. unwanted destruction or loss): for example ransomware attack;
- integrity breach (i.e. undesired alteration).

Whenever a data breach happens, the data controller has the obligation to notify the data protection authority within 72 hours of its acknowledging.
If there is a high risk for data subjects, the data controller must also notify them (Ho-mobile case, Italy, 2021).

Notification procedure - actors in the process

A lot of entities are involved in the data breach notification procedure. For example, marketing dept. is involved in order to write a message to users which do not frighten them. In 2007, Unicredit bank had a little security problem with personal data. They did a bad communication to their users, the result was that over 3000 people closed their bank account.
Key point: all entities must be involved because they could be useful, you have only 72 hours and you need to make good communication in order not to lose clients and not be fined by the data protection authority. Thus, the company needs to have an incident response plan ready to be instantiated in case of emergency, like evacuation plan in case of fire in a building.

Content of the notice

The notice must contain:

- nature of the personal data breach, including the categories and number of data subjects and the type and number of records involved;
- the contact details of the data protection officer (the authority communicate with company through this figure);
- describe the likely consequences of the data breach;
- describe the measures taken or to be taken in order to remedy the breach and mitigating risks for data subjects.

Key point: you cannot minimise what happened, you must describe your plan to fix the mess in order not to be fined by the authorities. Any kind of wrong assumption reported in the notice, can realise a false declaration, therefore a criminal offence.

Furthermore, the data controller must keep a **data breach registry**, which is an internal log to record all data breaches that happen in the company.

Risk management and DPIA

It is important to remember that the data protection impact assessment must be carried out since the beginning of the project, and must be a continuous process.
You have to take constantly into account the impact and risk of the personal data processing involved in your project.
DPIA cannot be done at the end of the project, since it affects the way in which you design the application.
Furthermore, you have to always think about the data minimization principle: only strictly necessary personal data must be processed.

Examples

Bike sharing:
Let us assume that a bike sharing company wants to perform sentiment analysis on their users while they use the service (i.e. ride a bike). Using a system for facial recognition to achieve this scope is not fine with respect to the GDPR, even if the company asks for the consent of the user. This is because facial recognition is an invasive tool, so we need to understand if there is another solution to perform the same task, but less invasive. For example, the company could propose a questionnaire at the end of the journey.
The same technology (sentiment analysis through facial recognition) can be used legally in another context. If we want to analyse the sentiment of a paralyzed person (i.e. cannot move and speak) it is legit to use this kind of system.
So, the key point is the following: we have to find a balance between the interest of the company and the interest of the user, and each case must be analysed on its own, understanding the whole context.

Airport:
In order to guarantee security on aeroplanes, all passengers have to pass through a body scan system. We need to find a balance between privacy and security. Since the body scan is an invasive procedure (with respect to privacy), security officers must provide an alternative solution for those passengers who do not want to pass through a body scan system.

Fingerprint:
A company wants to provide access to some specific areas only to authorised employees. In order to identify an authorised employee, they want to use fingerprints. This is an invasive data processing since fingerprints can identify a natural person and cannot be changed (a person can change its phone number, e-mail address, but not a fingerprint or any body feature). Then, they need to provide a different solution for those who do not want to provide their fingerprint data.

Formal assessment

When there is a high risk (impact) to any natural person, directly or indirectly derived from a personal data processing made by a specific company, the data controller must perform a formal assessment, assessed with some specific criteria (SEE SLIDE PRESUMPTIONS). Some presumptions are reported in the article

15, paragraph 3, of GDPR, but this is an open list (i.e. not exhaustive). In fact, article 35 of GDPR allows data protection authority to add cases of high risk (common law system approach).
Note that also this assessment must be carried out with a context based approach. For example, the large-scale notion is strongly dependent by the context of application: if we perform the processing in a small village is different than in a big town (since we have different orders of magnitude in terms of amount of data subjects involved).

Case study: wearable robots, exoskeletons
A company wants to use exoskeletons in order to reduce the stress and the effort made by its employees. They want to create a personalised exoskeleton for each person, so they need to collect quasi-biometric data (e.g. body shape, height, weight) about each employee. As mentioned above, quasi-biometric data are a particular kind of data, since a natural person cannot easily modify them. Thus, this system is not GDPR compliant.
The same system is GDPR compliant if it is used with a person with reduced mobility, in order to improve his/her life.
Back to the company case, instead of using a full body scan of their employees, they can produce exoskeletons which fit a range of people, for example basing on a height and/or weight range.

Case study: self-adapting workstation
A company wants to provide its employees self-adapting workstations, which can change some characteristics with respect to some biological parameters of the workers (e.g. stress level, etc…). The main issues are the following:

- the company collects real-time users' biological parameters;
- the company, with those data, could perform productivity monitoring;
- collected data are related to a natural person (the employee), then the company needs to face data protection issues;
- even if the company does not connect data to the employee, but only to the workstation, it can indirectly obtain the identity information since it has a work shift table.

Key point: How long and where does the company store that information? In order to make a GDPR compliant system, data should remain locally in the workstation, and deleted right after the employee's work shift.
Note that in working contexts it is not easy to demonstrate the users' consent for data processing, since there could be conflicts of interest. Thus, a company should not rely on that legal basis.

Large-scale project management
Previous case studies were specific to small environments. In large scale projects the situation is different (they are more complicated), for example because people often do not have the option to avoid the personal data processing.
In order to be compliant with GDPR, in a large-scale project only anonymous data should be processed.
Note that "anonymous" means that in every moment data are not related to any natural person. If data is initially connected to a natural person, then it is "anonymized" data, and until they are totally uncorrelated with the data subject they fall within the GDPR scope. Thus, you have to think about data protection.

Data management plan
It is a useful tool when facing a large-scale project. Note that it is not mandatory and it is not required by the GDPR or any other law.
It maps the data flow and who is involved. It defines a uniform way of working among various subjects involved: this is necessary in order to avoid a mess.

Large-scale case study: smart mobility
Mobility data is sensitive data since you can infer political or religious information about the subject, for example if he/she goes to a church or something else.
In 2014, London's public bike sharing service decided to publish their usage data. The main issue was that data were only pseudonymized, since they were connected to a user's identifier. The identifier alone cannot be associated with a natural person, but with all usage information of a single person, in some cases it is possible to identify that person (for example looking at the most frequent trips).
Then the solution is to completely remove the IDs column, so that journeys are not grouped by means of a natural person usage.
If the company needs to associate the usage with a natural person, then it must be justified, and data (about ID) can be retained only for the strictly necessary time. Of course that kind of data cannot be published.

Another interesting case is Piedmont's smart public transport systems. GTT and all the other public transport companies in this region use NFC tags as tickets, in order to analyse usage of the public transport. Since usage is related to a natural person (even if you cannot identify him/her using only the ticket number), companies need to think about data protection. If companies keep that data in a pseudonymized form, they need to declare their intentions. Otherwise they must remove any link to a natural person.

Case study: enschede Wi-FI system
A municipality in The Netherlands decided to create a free Wi-Fi hotspot system in a city centre. Their aim was to collect data about the crowding of some areas, looking at the number of phones connected to each device of the system. The main issue of this system was that they also associated an identifier to the device, then the system was actually a tracking system. This directly follows by the fact that they didn't follow the "privacy by design" approach.
They were sanctioned by the data protection authority even if they did not use the collected data to track people, just for the presence of the risk of tracking.
Key point: data protection regulation is a risk-based law, and its aim is to create a safe environment for users.

Regulatory competition and global data protection
All countries are interested in expanding their data protection regulation all over the world. Nowadays, GDPR is the golden standard, and works well (only on paper). It does not work well in real life since it is not possible to impose a company outside the EU to have an establishment in it.
Key point: the current data protection regulations are not enough for artificial intelligence systems.

Beyond data protection - AI context

Case study: ClearView AI
The Swedish police used clearview AI (based on facial recognition) to process personal data, in order to identify suspected people. The main issue was that the processing operations were carried out without performing an impact assessment before starting.
You cannot use facial recognition for each criminal case, since it is a very invasive technique.
The problem, again, is proportionality: you need to assess the risk for each situation, specify your purposes and demonstrate that the balance of interests allows you to proceed.

Different scenario: research project
A research group in the UK decided to create a dataset composed of images collected from social networks. After that, they wanted to make the data publicly available. The problem was that, among those images, there were photos of people (also minors). Since users did not express a consent for this kind of processing for their contents, there was not a valid legal basis to process those personal data. Even the research legitimate interest is not enough, since the privacy legitimate interests of users prevails.

Case study: progressive bias
A car insurance company decided to force their users to install a black box system to record their car usage, in order to obtain the best price on their insurance policy.
Among the parameters used to establish the price, there was the amount of time in which the car was used at night, since a lot of people return home from parties, often drunk, at this moment of the day.
The problem is that they didn't consider who worked at night. So there was a category of people penalised by this system without any valid reason.

Two dimensions of consent

The consent of the data subject, in the context of data protection, is a system to negotiate your personal data in exchange for free services (for example). Theoretically, each user decides whether to provide his/her personal data or not, doing a self-assessment of the risk. This is the aim of the informed consent, i.e. to provide to the user a picture of risks deriving from a certain data processing.

What are the impacts of AI in this scenario?

- *Datafication*: everything is transformed into data, since they are collected from every electronic device. Then, we generate data in each moment of our life (both personal or not).
- *Control over information not only silos, but also open data:* the aim of open data (like those of the public bike sharing system in London) is to reduce the gap between companies with a lot of data and those without. But, instead, they increase this gap because only some entities have enough resources (computational and human, i.e. data scientists) to extract knowledge from those data.
- *Data portability*: it works well only if this right is ensured also in a technical environment, not only theoretically. For example, if a company wants to change cloud provider (e.g. migrate from AWS to MS Azure) there is a cost for training employees. Again, if a person wants to migrate from Whatsapp to Signal, there will not be the same people. Then portability is not only a legal issue.
- *Categorical approach*: before AI, categories of people were static and they were made with respect to some basic characteristics like religion, political thoughts, race, etc… This allows a specific group of people to get in touch with other subjects in the same cluster, and fight together for their rights in case of a discrimination or some other threats. Nowadays categories are inferred from data, and are no more static. This is an issue for multiple reasons:
    - first of all, we do not know which other subjects are in our same category, therefore if a group of people is victim of discrimination they cannot fight for their rights;
    - the system is opaque for the user, he/she doesn't know why he/she is classified in a certain manner;
    - the categories are dynamic, i.e. the list of categories is strongly dependent on data and a user can be reclassified, since the classification is not done only on permanent characteristics of the person.
- *Purpose specification principle:* the purpose of personal data processing should be presented at the subject in the moment in which data are collected. In AI systems, data are often used in the support decision making process. It is not known a priori which type of information AI will infer from data. For example, from mobility data it is clear that information about the traffic can be inferred. But from this kind of data can be inferred also clusters of drivers depending on their driving behaviours, the presence of events, etc …

another example is university, which collect the info of student by the law and they won't ask for consent of students while they are collecting info of students.

While in a hospital, with health data, an AI can infer also the (health) habits of a person and not only the information about the disease which the doctor is looking for.
Key point: if you do not know what kind of information you will extract from data, you cannot specify it at the moment of collection. Thus, there is a conflict with the GDPR principles.
Different solutions were proposed.

- *Dynamic consent*, i.e. asks for an update of the consent to the data subjects whenever the purpose of the data processing changes. This is not a good solution since it has negative repercussions on the behaviour of the users: they will accept or reject every request without even reading it.
- *Broad consent*: if you have only a general idea of the purpose, you can use that. Be aware that it could be against the principle of purpose limitation.

general consent means we know general purposes not exactly

Trouble of risk assessment

The DPIA focuses only on issues about the processing operations (e.g. cybersecurity, etc…), but not on individual rights issues (e.g. discrimination, freedom of expression, etc…). Furthermore, a lot of companies underestimate the risks of data processing, and the data protection authority does not have enough resources to check everything. Note that the manipulation of the DPIA, done by the company to proceed with a data processing even with a high risk, has not a big impact on the sanction in case of realisation of risk.
Thus, we can conclude that data protection is not enough for AI applications.

AI vs human intelligence

In order to emanate a good regulation, we need to understand the context. Nowadays, AI is based on a data-centric approach, focused on mathematics and statistics. Let's see a concrete case. Amazon decided to use AI to analyse the curriculum vitae of candidates applying for some open positions. They used curriculums of employee to train the model, but removing the sex information in order not to have sex bias. The result was that the model was biased in selecting males against females. This was due to the fact that the sex feature was strictly correlated with some other features that were not removed from the training set.
Key point: machine reasoning is different from human reasoning, and we need to be aware of possible biases.

the AI reasoning is different to human(can have more bias and is our mood can have an affection on it when the AI having bias all of decision will have bias but for human it could be different due to his mood) and it's based on historical data and statistical info. When the result of AI and HI is same we can not say the reasoning is same

How to define AI?

Law needs to provide definitions of what it is regulating. Otherwise, if a case does not fall within that definition, the law is not applicable. Therefore there is a risk in defining what AI is, since it changes rapidly over time.
Scenario of legal person: a legal person is for example a company, it can act like a natural person (e.g. signing contracts, etc…) but it doesn't exist in reality, only for the law, and this is the main issue of the legal persons system.
A proposal was to represent AI as a legal person, but it is too simplistic and, in the end, it introduces a new layer of complexity without solving the problem. Then this proposal was rejected.

Transparency in AI systems

This concept is not well defined in an AI system. If we know that a certain variable influences the result but we do not know why, it is quite useless. Furthermore, most of the time the user is not interested in transparency, in fact few people read the privacy policy when they use a service. Then why should users be interested in transparency when using an AI system?
An example of this concept is summarised by the case of COMPAS application, used in the USA to predict whether a prisoner could be released before the end of the sentence. The problem was that this application was biased by the colour of the skin, and tended to classify as "not suitable" black people.

This is why we need transparency in AI systems, it is needed to perform a proper risk assessment.

What are the risks of AI?

- Risks of bias: not only in the dataset, also in methodology, like the case in which the insurance company classifies as “bad drivers” those who use the car at night. It is also in the source of data (*selection bias*). Furthermore, a bias is introduced by data scientists, who shape the product with respect to their point of view (*confirmation bias*)
- Decontextualization: an AI system often becomes unreliable when used in a different context than the usual one. Then we should consider this fact in the risk assessment.

Do we need a new regulation for AI?

Yes, data protection regulations are not enough, but we do not need to reinvent the wheel. We need to update the current regulations in order to fill the gap with AI technologies. We have issues affecting groups (see above where we talk about the dynamic classification issues), not individuals. Furthermore, AI is a peculiar product: it learns from the users, then who has the responsibility? The user, the producer or the product?
How to regulate? with hard laws (e.g. regulations, etc…) or soft laws (e.g. code of conduct, etc…)? The first kind of laws has a top-down approach and are more effective, while the second type has a bottom-up approach and are less effective.
Key point: the current regulations do not protect groups of people, only some individual rights.

Key issues in regulating AI

Whenever there are new kinds of issues, in terms of regulation, there are many different solutions. Let us present them.

- *Interpretation*: when people have a problem, they go to the court. Then, the court is the first entity which faces new issues. Therefore it is the first actor in emanating a new law, also before than the legislators. The problem is that the court needs to manage each case (if the topic falls within the scope of the law) even if there are no specific laws for that case. For example, when the Internet became popular, there was an issue in defining what is an ISP (Internet Service Provider) and whether it is liable for the content online (e.g., infringement of copyright, prohibited content, etc.). The law allows one to choose among **editors** or **publishers**. The first are liable for the content, the second are not. The resolution of this case was a paradox, since the ISP was defined as an editor if it tried to block some kind of content, while if it didn’t apply any sort of operation to block contents it was classified as a publisher. Another problem of this approach is that the decision is weak, and it cannot be easily applied in a future similar case. Furthermore, courts are not in the position to fix the law, in the ISP case they had to choose among editors or publishers, they couldn’t create a third category.
- *New regulation*: for example the GDPR, which introduced the concept of risk-assessment based data protection, is a new regulation. Before GDPR the situation was completely different.
- *Extension of existing regulations*: this happened for e-commerce. There was already a law for physical stores, the main issue was that the client could not see the object before buying it, and he/she was not sure that the store was still there the next day. Therefore, it was only necessary to extend existing regulations for normal stores in order to manage those specific cases.

Then, in which way should we regulate AI? The only solution is to introduce a new piece of law, but carefully, because it must be consistent with other laws already existing (in data protection context there are a lot of other regulations). For example, it was proposed to create a European authority specific for AI: it was rejected since this kind of cases fall within the scope of data protection, therefore there would be a conflict with Data Protection Authorities.

Another issue in regulating AI systems is that problems depend on the context in which AI is used. For example, if it is used to support a decision making process, there is the supervising issue, which is not present in other contexts. Let us explain this issue: a bank officer must decide whether to grant or not a loan. Based only on personal experience, the officer would provide the loan, but the AI used to support the decision returns a negative evaluation of the applicant. If the loan is provided overriding the AI decision, in case the borrower will not pay, the officer will be liable. Then, if the AI is biased from data used for training, there is a problem.

So there are two possible ways to emanate a new piece of law.

- *Co-regulation*: we mix a single hard-law (i.e., a regulation) with a soft-law (i.e., code of conduct) for each specific sector of application. An infringement of a soft-law would result in infringement of the hard-law, therefore a sanction. This happens in the GDPR.
- A hard-law for each sector of application. The professor seems to prefer the other option.

SEMINAR: Digital Mare Nostrum - the Jean Monnet Chair

More here (credits: Francesco Capuano)

The Jean Monnet Programme, also known as the Jean Monnet Project or Jean Monnet Actions, is a European Union initiative to encourage teaching, research and reflection in the field of European integration studies in higher education institutions. The Chair aims to promote Mediterranean studies in computer law through an interdisciplinary cross-fertilisation, contributing to reduce the existing gap in the field of computer law that affects Mediterranean countries, which are often underrepresented in the global arena. The Chair wants to amplify the voice and participation of Mediterranean countries in the EU regulatory and policy debate. This will enrich the discussion at EU level through a focus on digital societies, benefiting from the significant scientific output in humanities and legal studies of Mediterranean countries with a positive interdisciplinary cross-fertilisation for the development of human-centric digital societies and AI.

Keynote

This keynote was taken by a member of Homo Digitalis, a humanitarian association which focuses on protecting digital rights of refugees in Greece. The association is the result of a fusion between two entities.

- The first focuses on digital rights, data protection issues and biometric recognition.
- The second is about social justice, like social pushbacks, non discrimination, arbitrary detention lack of process, access to health housing.

Some real cases in which Homo digitalis was involved are:

1. Social media analysis and monitoring to predict illegal migrational flows (cancelled plan).
2. Monitoring instant messages (e.g. Telegram) in order to predict migrational flows.
3. Facial recognition for mass surveillance, used to detect irregular immigrants (blocked by the italian data protection authority).
4. Behaviour surveillance system used for security in migratory facilities. Used to identify bad behaviours and send drones. They went against this system but they failed, then a coalition was created and they used article 57 of GDPR, now the case is being evaluated by the greek DPA.

Key point: the GDPR is applicable not only to EU citizens, but to all the people in the Union (including immigrants).

Panel

The first intervention in this panel was made by the head of the data protection office of the international committee of the Red Cross. He talked about data protection in the humanitarian sector. The key point of this speech was the difference between the approach of big companies and humanitarian associations to data protection: the first ones can decide to not operate in a certain country, specialise their procedures to

satisfy the data protection regulation or face sanctions; while the second needs to operate in certain countries, they cannot decide to avoid certain country for data protection issues. Furthermore, all the national sections of Red Cross are independent, therefore the issue is to find common procedures in order to respect the data protection regulations of all the countries in which they operate.

The second intervention was about how AI systems are used to track people crossing borders. It can both improve logistics and discrimination against some people based on ethnicity, it depends on how it is used by governments. Usually, these kinds of new tools do not have a legal basis against the rights they infringe. Furthermore, some systems are not mature, with lack of scientific evidence.
The solutions to this problem are:

- Follow a risk-based approach
- Be sure to respect fundamental principle
- Declare which AI applications can never be used
- Introduce transparency in AI systems
- Right to choose interaction with humans instead of machine
- Focus on preserving democracy in case of elections.

The third intervention was about the interaction of humanitarian associations and the big tech companies. Often, humanitarian associations lack resources, because of their non-lucrative nature. Then, they need to interact with private companies. The latter are useful to exploit data, and they offer their services for free. For example, the World Food Programme collaborated with a private company to collect and process data on their activities. The company could not use data for other purposes, but the aim of the company was to be protected against the law even if their behaviour in other contexts was not very transparent, since they were associated with a very important humanitarian action. The same happens with Meta, which wanted to create a crypto currency called Libra to be used in humanitarian action, and with Whatsapp which is being used in the current Ukrainian crisis. The primary aim of Meta is to be protected against Antitrust, not to help people in Ukraine. Another aim of these companies is the technological lock-in of the humanitarian associations, i.e., they cannot easily change data processors.

Roundtable

NB: THIS PARAGRAPH IS NOT COMPLETE AND SOMEONE SHOULD ADD NOTES
This part of the seminar was about how the EU uses AI on immigrant people from a law point of view. Borders are human labs for AI experimentation. The EU actively supports such programs, since immigrants are not represented in the parliament. In these kinds of activities, data protection is not considered, in fact a lot of databases, created for different purposes and correlated to immigrant people, were merged in order to have big data, blurring the purpose. For example, in soup kitchens, ear scanning is used to detect whether a person has already eaten (this is biometric data, which is like face recognition).
Furthermore, AI is being used to classify travellers according to several risk profiles, in order to prevent immigration risks and public health threats. It is also used to farthen EU borders in Turkey: EU supports the Turkish government in the use of AI systems to detect migratory flows across the country and directed to Europe.

Ethics in AI regulations

At the moment, there exists an European commission composed of high-level expert groups for regulating AI systems. The main problem of this commission is that it is not independent: in fact, the majority of experts come from big tech companies and governments. The few academic researchers present are supported economically by private companies. Their workflow is the following: they divide the committee in subgroups, each one is debating on some AI-related topic; at the end they merge results.

This committee considers data ethics a good solution for AI-regulation issues, but there are many conceptual problems.
First of all, they use the notion of “trusting in AI”, which is not well defined: it is an irrational concept, since the “trust” is based on relationships (when it is considered between people), so it depends on how the product is seen. Then, a good marketing strategy could mislead a lot of users. **Trust is a tricky concept,** you can think about media, which can be easily manipulated with fake news: why do you trust a specific media? Because of his name? Because of your bias against the subject?
Let us say that we trust in an AI system because the producer strictly follows an ethical framework. There is no ethics based on human rights, but human rights based on ethics. Therefore, it makes no sense to state that an AI system should be based on an ethical framework in order to grant human rights.
The commission mentioned above states exactly this in their document, therefore the produced document is confusing, it is difficult to understand what is needed for an AI-producer company.
Furthermore, they create ethical guidance related to legal contents, then there is a conflict. For example, they claim the accountability principle, which is required from GDPR when dealing with personal data. Then a company could think that this principle is only an ethical recommendation, while instead it is mandatory.
The same is transparency, which is required by some other laws and it is an ethical recommendation of this committee.

The Seven Key Requirements for Trustworthy AI

1. Human agency and oversight
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non-discrimination and fairness
6. Environmental and societal well-being
7. Accountability

have natural act, reasonable by human

tolerating due to anything which may affect on system

Another problem of these documents is that they do not cite any philosophical source, they contain only a list of principles: how do they conveyed the usefulness of this list? Furthermore, a philosophical framework is present only in some parts of the world, therefore you cannot imagine how to regulate AI all over the world without considering all the philosophical thinking from different countries.
Another key point is that ethics and laws must be on different planes. There are two ways to do that:

- With a *complementary approach*, i.e., you start from an ethical framework, you define the main principles and then emanate a law to express how companies must implement those principles in a practical manner.
- *Reconsidering existing laws*, i.e., you discuss basing on an ethical framework and then choose among existing laws.

Keypoint: Nowadays, there are only principles and recommendations for regulating AI, which are useless for companies which produce AI. There are many different ways to implement a principle, therefore laws must clarify the implementation issues.

Towards an European approach

Nowadays, the maturity of the European proposal (for regulating AI) is the most advanced in the world. It sticks a white paper approach based on hard-laws, therefore it follows a top-down approach. It revises existing legal frameworks, like the GDPR. This is not so good since this last was revised very recently. Furthermore, you should engage different people in regulating AI, you cannot use a strict top-down approach.
Another open issue is: how should we define AI?

A current proposal is to define it based on what is the specific task of a system. This point is crucial, since AI techniques evolve very rapidly over the time, impacting society in many ways. Then, a broader definition should be provided.

Another issue is how this proposal defines *high risk* of an AI system. At the moment, the proposed regulation applies only to high risk applications, but a lot of low risk systems can become high risk. Furthermore, when technologies are in an early stage, it is better to be not so restrictive with regulations since it could slow down investments in that sector.
The high risk definition proposed in the regulation is based on the impact on society. The main issue is that the "high risk" concept is defined in the text of the law (some technologies, regardless the context, are considered risky), unlike the GDPR which proposes a risk assessment case by case. Having a close list to define high risk is quite dangerous: in 10 years the situation will be very different.
Another way to define high risk could be with a sector approach, i.e. define it with respect to the application sector. For example, if an AI is used to control a huge machine then it is high risk, regardless of the technology used for developing AI.Note that the high risk is acceptable if it is mitigated with some appropriate measures.
Although, the impact on human rights is not well regulated in terms of risk. It is quite difficult to assess the impact on human rights of a technology: it can be of a very good quality but impact negatively on human rights.
The proposed regulation divides AI systems in three categories:

- Safety law I
- Safety law II
- Stand-alone AI

The first two are already covered by other regulations in terms of quality, while the last is all new.
Keynote: Note that in the proposed regulation, there is no assessment to explain why some technologies are considered high risk or not. The commission is not transparent, there are no objective criteria specified to explain their choices.

What is missing?

- *Flexibility*: this concept refers to the fact that there is no impact assessment for evaluating an AI system, this is because it would stop investments.
- *Open clause*: defining a close list for classifying an application as "high risk" will result in companies that try to stretch the law in order to make their application falling out of the scope of the regulation.
- *Decentralized approach*: there is only a commission to emanate this regulation, without a variety of experts from different sectors.

Furthermore, there is not a model to perform a human rights impact assessment for AI systems: we should extend the data protection impact assessment because it includes human rights, but it does not consider the context.

Case study: Hello Barbie

The product was a doll which could provide an answer to kids' questions. Here we report the main issues:

- Right of education: the answer of the doll could be fine for some cultures, bad for others.
- Advertising
- Privacy and data protection: data are collected during the dialogue with the kid.

Then an impact assessment is necessary before deploying this kind of product on the market, since it helps to take the right decision. For example, when you have to decide how to retrieve the answer you have two options:

- Pick the answer from an open list (online): more risks

- Pick the answer from a closed list (e.g. a database), maybe personalizable by the parents: less risks.

Conclusion of data protection part

Problems in creating a code of conduct: منشور اخلاقی

- It refers to general principles
- It is difficult to be implemented and there are a lot of ways to interpret it
- There is no evidence about the level of compliance
- It could be used as a marketing tool

When new technologies come in, new ethical issues arise. With AI you see the world through this new technology, therefore there is a bias induced by it: it can give a new shape to things and change the perspective. An example is the content moderation on social platforms, which provides a different truth. Sometimes, this is unintentional.

Data ethics

Data ethics is useful to build devices and digital infrastructures that are not discriminatory for the weakest part of the society.
Ethics, also called moral philosophy, is the discipline concerned with what is morally good and bad and morally right and wrong. The term is also applied to any system or theory of moral values or principles.
Its subject consists of the fundamental issues of practical (data-driven) decision making.
In this course we have a focus on certain aspects of ethics:

- Epistemic concerns: the data we are using is not good enough, the data is insufficient or we cannot look at the data used to build the system.
    - Inconclusive evidence
    - Inscrutable evidence
    - Misguided evidence
- Outcome of a system is then used to trigger and motivate an action that may not be ethically neutral.
    - Unfair outcomes
    - Transformative effects: how the inequality in our society is changing when we automate processes which affect life of people
- Traceability: complicated apportionment of responsibility for effects of actions driven by algorithms

Introduction to data ethics

Nowadays software is programming the world, it is shaping our society.
Automated Decision Making (ADM) systems are used, for example, to identify the best suitable candidates for a job position, to take content down from a social network, to detect social welfare frauds, to predict the risk of social violence at home, to suggest which university to attend, etc.
ADM systems learn from historical series (*inductive learning*) or examples and make decisions. In order for the software to learn correctly the following things are needed:

- A sufficiently large number of examples
- A sufficiently heterogeneous set of examples
- Examples annotated with the right answers (who is doing this annotation?)

There are different types of inferences (learning methods):

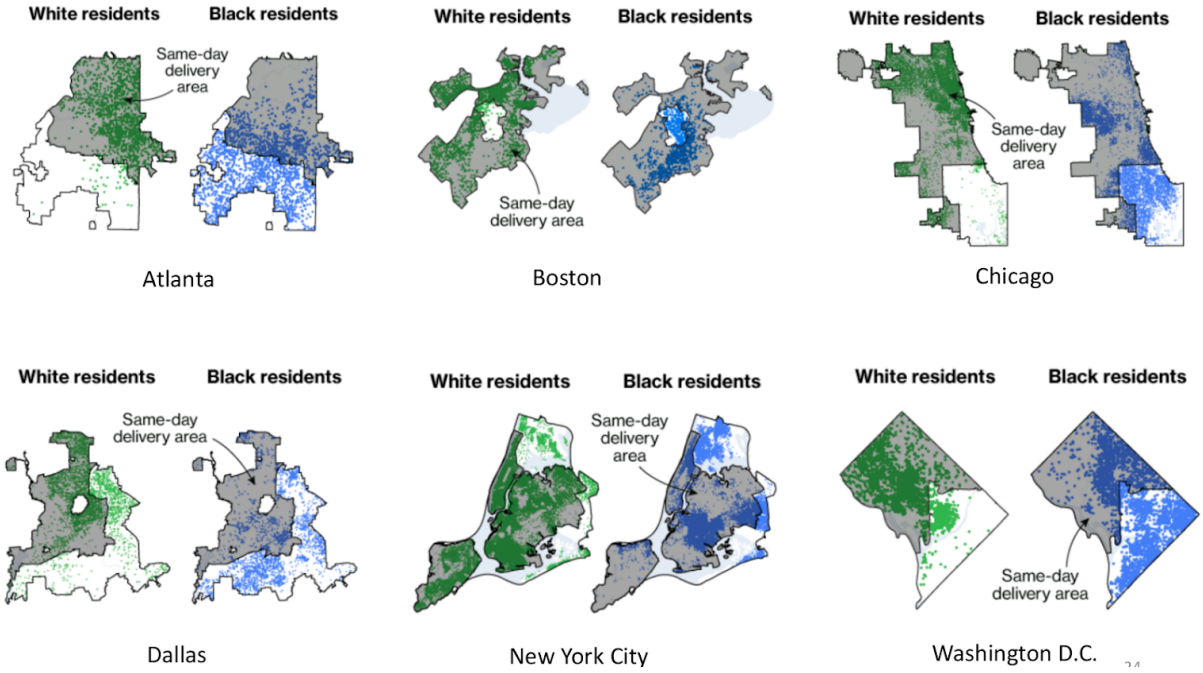
- Inductive: 96% of Flemish college students speak both Dutch and French. Louise is a Flemish college student. Hence, Louise speaks both Dutch and French.
- Deductive:All As are Bs, a is an A, hence a is B.
- Abductive: I observed many gray elephants and no non-gray ones. *The best explanation* is that all elephants are gray. I infer from this that all elephants are gray.

Machine learning is an inductive process, since it starts from training examples, defines statistical patterns and makes inferences on new instances.
There are some problems with this approach:

- Reality is a super-set of what is measurable.
- Some aspects of our society are measurable only indirectly
- Societies have historical and structural inequalities, reflected by the data
- Spurious correlations and confusing factors: we can always find correlations in data.

Case study: demographic disparities

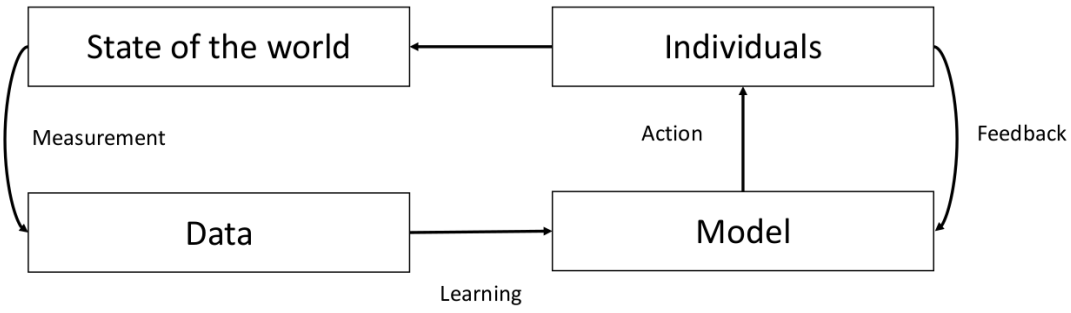
Amazon decided to analyze historical data of online purchases on its platform and demographic data to determine in which neighborhoods to offer the fast (same day) delivery service. This happened in the major cities of the USA.



As you can see from the figure above, the service was delivered where the majority of the people is white. 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. Neither the source code nor the logic of the main algorithm is inspectable (*inscrutable evidence*).
Open discussion points:

- Should the source code or the logic of the algorithm be inspectable?
- Should data used by the software be anonymized and made public?

Demographic disparities in the loop



State of the world
When looking at the actual state of the world, disproportion and inequalities are common in our society. For example, looking at the city of Turin, we have unbalanced distribution of age, income and sex with respect to the district considered (e.g. higher incomes are in Aurora and San Salvario).
In the United States, a big gap is in the income of people w.r.t. their ethnicity. In the first slide of the lecture, many cases of unbalanced data were presented.
Keypoint: unbalanced data is the normality, but we have to be aware when we measure the reality with data.

In fact, data about society and individuals will reflect these disparities, because during the measurement process many subjective choices have to be made. This process consists in assigning numerical values to an entity, with the purpose of characterizing a specific attribute. Let's see an example.

Measurement (Data)

EXAM QUESTION: what are the measurement problems?

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 problem with these measures are many:

- Some programming languages are more verbose than others (the measure should be weighted by the programming language)
- The indicator does not take into account the time spent on other required activities (e.g. producing documentation, requirements, etc)
- Moving from a rational to an ordinal scale eliminates distances between positions (i.e., the gap between the first and the second can be much higher than the gap between the second and the third)
- People will change their behaviour in order to get a higher score (for example a developer could start writing verbose code to do the same operations, which is not a good thing in SWE). This behavior is also observed in the research world, where researchers need to have a high H-index to cover better positions.
  Therefore, the more an indicator is used, the more it will be subject to the pressures of corruption (*Campbell's law* or *reflexivity problem*)
- When a measure becomes a target, it ceases to be a good measure (*Goodhart's law*)

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 uses defects fixed over worked hours (fixed defects per hour)
- The second uses the reciprocal, worked hours over defects fixed (hours for fix a defect)

Data are the following:

- Group 1 (G1) contains Developer 1 (D1) and D2, G2 contains D3 and D4.
- D1 fixes 1 defects per hour, therefore it takes 1 hour to fix a defect (reciprocal)
- D2 fixes 4 defects per hour, therefore it takes 0.25 (¼), a quarter of an hour to fix a defect.
- D3 fixes 2 defects per hour, therefore it takes 0.5 (½), a half of an hour to fix a defect
- D4 is the same as D3.
- The average for group 1 is 2.5 fixes defects per hour and 0.625 hours to fix a defect
- The average for group 2 is 2 fixes defects per hour and 0.5 hours to fix a defect.

The first manager will declare group 1 as the best because they fix more defects in an hour, while the other manager will declare group 2 as the best because they take less time to fix a defect. Note that they both used the same data. This is known as *Simpson's paradox*, and it is one of the most common errors in decision making systems. It is due to the non-linearity of the reciprocal operation.

Measuring involves defining the variables of interest, defining the process to interact with the real world and transforming the observations in numbers (like in the previous example). Usually software developers do not follow the whole process, but they use data according to some given requirements.
The problem is that data quality might be not suitable for the intended use and the real world is much more complex than variable-value pairs, and the rules defined to analyze them.

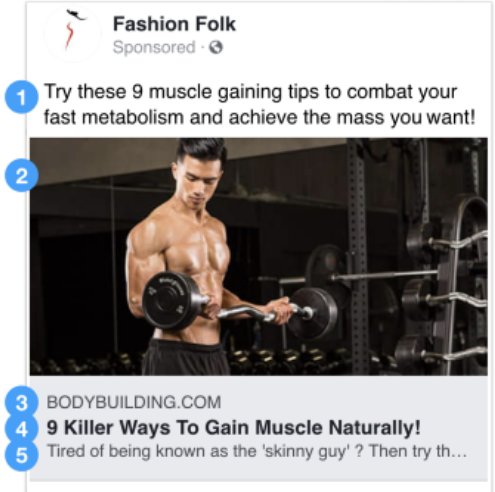
Learning (Model)
Models can propagate demographic disparities present in the data. For example, if the admission models to American universities had been trained on the basis of data from the 1960s, we would probably now have very few women enrolled, because the models would have been trained to recognize successful white males.
Another example of a biased model is Google Translator, because it assumes “doctor” as male, “nurse” as female.

Action (Individual)
The model is trained on data from the entire society and then it is used to predict characteristics of individuals. This predictive policy is the main issue:

- The characteristics and behavior of individuals change over the time
- The output of an algorithm can have an effect on the individual, which tends to confirm or contrast the action (Campbell’s Law and Goodhart’s law)

Furthermore, only certain behaviours can be mapped, some others are unpredictable.
(Many other examples are reported in the slides).

Case study: the Facebook advertising platform
Introduction
In this case study we will present the issues related to this platform and the legal action against Facebook.



An ads is composed by:

1. Headline and text
2. Images and/or videos
3. Domain (pulled automatically)
4. Title (pulled automatically)
5. Description (pulled automatically)

The advertising process is composed by these phases:

1. Ad creation (ad contents, audience selection, bidding strategy): this is managed partially by the user
2. Ad delivery: this is totally managed by Facebook

In the performed study these effects have been evaluated:

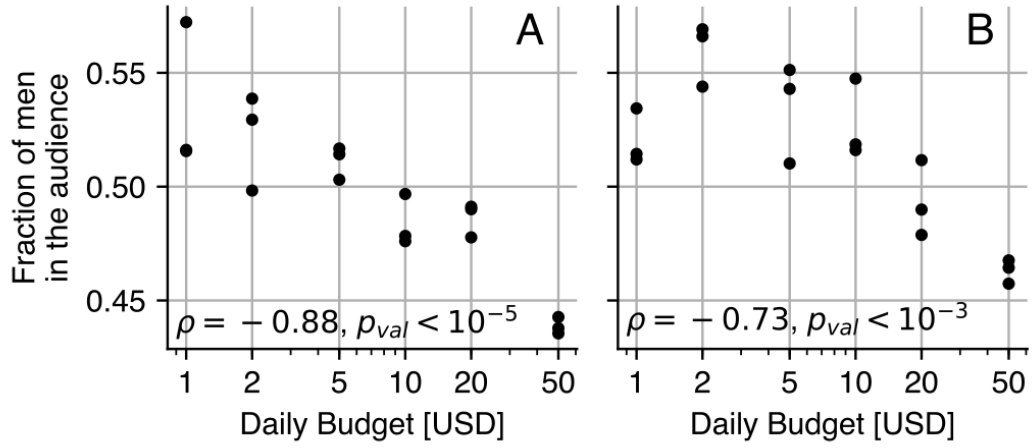
- Budget
- Ad creative
- Ad image
- Ad image classification
- Test on real-world ads

Researchers investigated whether the platform integrates an advertisement and how. It was discovered that Facebook performs a labelling operation on images contained in the ads. The Facebook defence during the

legal action was based on the statement that the platform is neutral, but this is not true as it will be shown. In fact, during the ad delivery phase, Facebook is fully liable since it decides the audience on its own. During the study, four regions in North Carolina were taken into account. Two of them were composed of approximately a half million of white people each, while the other two were populated by the same number of black people. Then the first two are targeted as “white” since the majority of people is white. The same for the last two ones. This is a simplification of the study; furthermore, samples are comparable in terms of quantity.

Budget effects

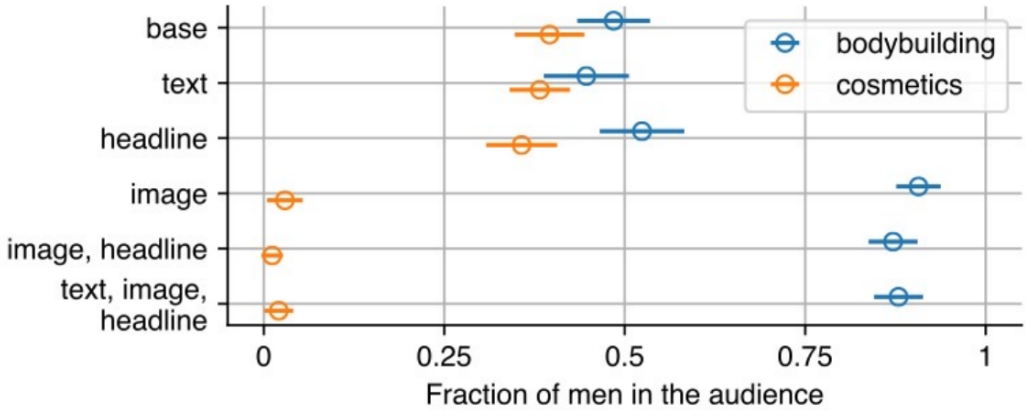
The first experiment aims to measure the percentage of men in the audience with respect to the budget level involved to create the ads. Here is shown a plot of the results.



In plot A were considered all users located in the U.S., while in plot B a random custom audience. As you can see, when the budget increases there is a lower fraction of men in the audience. We do not know the explanation of this behaviour.

Ad creative effects

In this experiment it was measured the percentage of men in the audience changing some features in the ad.



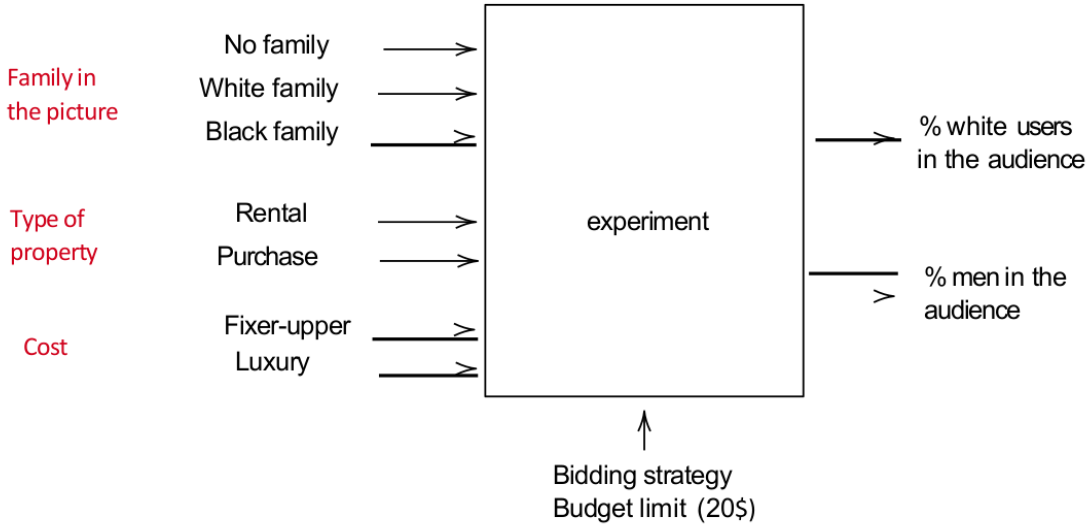
Two domains were considered: bodybuilding and cosmetics. The budget was fixed in order to not have wrong results. As you can see from the previous image, what makes a big difference in the sex distribution is the image. This result is quite important because it demonstrates that there is something going on with

the images in the advertisement: there is an intervention of the platform, because the content creator does not explicitly provide any information about that image.
Another experiment related to the one just presented, is using an invisible image and a normal one. The invisible image has the same structure of the normal one, but it uses a color palette invisible to the human eye. The same behaviour of the platform was noticed, and this proves the use of a classification system based on AI.

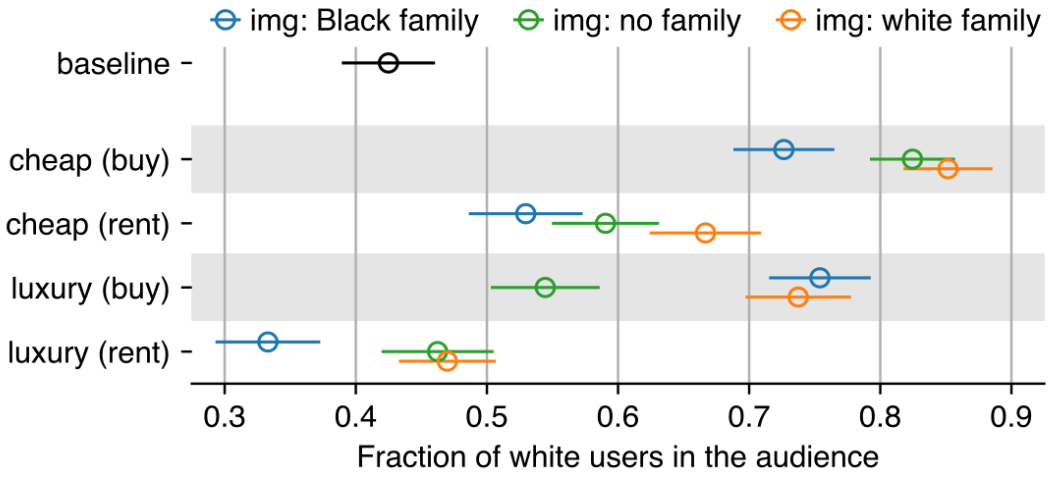
Other experiments

They measured the percentage of white people in the audience against the music type involved in the video contained in the ads. With country music, around 80% of the audience was white, while this number dropped down to 12% for hip-hop.
The same measure was tested with respect to the subject of the image in different ads for job offers. Guess what? Again discriminations.

The final experiment, the one which caused the legal action that will be presented in the next section, was about housing ads. The input parameter and the measures registered are reported in the bottom image.



And the results are here.



Housing and Urban Development against Facebook inc.

In the following, Facebook inc. will be addressed also with the term “respondent”.

Legal authority

The above results demonstrated that a discriminatory housing practice has occurred. In fact, from the late '60s, it is unlawful to make unavailable or deny a dwelling to any person because of race, color, religion, sex, familial status, national origin or disability. It is unlawful to discriminate against any person in the terms, conditions. It is unlawful to make, print, or publish, or cause to be made, printed, or published, any notice, statement, or advertisement with respect to the sale or rental of a dwelling that indicates any preference, limitation, or discrimination.

So, the key point is that it is illegal to show a housing advertisement only to some people.

Factual allegations

Respondent collects millions of data points about its users, draws inferences about each user based on this data, and then charges advertisers for the ability to micro target ads to users based on Respondent's inferences about them.

Because of the way Respondent designed its advertising platform, ads for housing and housing-related services are shown to large audiences that are severely biased based on characteristics protected by the Act, such as audiences of tens of thousands of users that are nearly all men or nearly all women.

Respondent sells advertisers the ability to target advertisements to people who, according to Respondent's assessment of the data it collects, share certain personal attributes and/or are likely to respond to a particular ad. In fact, it was possible for the content creator to select the race of the audience.

Respondent determines which users will see an ad through a two-phase process.

- First, in the *ad targeting phase* the creator chooses some characteristics of the audience. This is an unlawful design choice made in the creation of the system. Furthermore, it was possible to select "best existing customers", which was labelled in this way based on which pages a user visits, which apps a user has, where a user goes during the day and the purchases a user makes on and offline.
- Second, in the *ad delivery phase*, Respondent selects the ad's "actual audience," meaning Respondent chooses which users will actually be shown the ad from among the pool of eligible users. Respondent bases this decision in large part on the inferences and predictions it draws about each user's likelihood to respond to an ad based on the data it has about that user. Note that in this phase Facebook is fully liable, since advertisers have not the power of choice.

Furthermore, Respondent charges advertisers different prices to show the same ad to different users. Respondent's ad delivery system prevents advertisers who want to reach a broad audience of users from doing so.

A local law of New York City: Automated Employment Decision Tools

This law was created to regulate the usage of automatic decision systems in selecting personnel of the companies in NYC. In order to use such systems, it is required to perform a "bias audit", i.e., an impartial evaluation of the system by an independent auditor. The aim is to assess the tool's disparate impact on persons.

It is unlawful for an employer or an employment agency to use an automated employment decision tool to screen a candidate or a employee for an employment decision unless:

- The tool has been the subject of a bias audit conducted no more than one year prior to the use of such tool.
- A summary of the results of the most recent bias audit of such tool as well as the distribution date of the tool to which such audit applies has been made publicly available on the website of the employer or employment agency prior to the use of such tool.

Formalization of algorithmic fairness

In this section we will provide some criteria to evaluate in order to declare a machine learning model "fair". We will use the following notation:

- R is the output of a (binary) classifier, the prediction
- Y is the (binary) target
- A is the sensitive attribute (e.g. sex, race, …)
- $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ are the data observations, each pair (x_i, y_i) is:
  - An instance vector $x_i \in \mathbb{R}^p$ where p is the number of features
  - A label $y_i \in \{0, 1\}$

The following (strong) simplifications will be assumed:

- Data are independent and identically distributed (i.i.d.)
- The data distribution does not change over the time

Independence

A model respects the indepence criteria if and only if R (the prediction) is independent from A (the sensitive feature). In probabilistic terms:

$P(R = 1|A = a) = P(R = 1|A = b)$ where $A \in \{a, b\}$.

More precisely, we want those two probabilities to not differ more than a threshold epsilon.

Let's see an example in which independence is not respected. Here it is the confusion matrix.

Gender (A)	R=1	R=0	Y=1	Y=0	Tot
Female	6	4	7	3	10
Male	7	3	7	3	10
Tot	13	7	14	6	20

- P(R=1 | A=female) = 6/10 = 0.6
- P(R=1 | A=male) = 7/10 = 0.7

Since those two probabilities are not the same, the model which produced those predictions is not fair with respect to the sex attribute. In this case, there is a difference of 10%: in the real world you need to find a trade-off between accuracy (or any other quality measure) and independence. If you aim for perfect independence, then probably it will result in bad predictions. You can see an example in which independence is respected in the slides.

The advantage of the independence indicator is that you can apply it at every stage of the process, but on cons, it ignores the possible correlation between the label and the sensitive attribute, resulting in a worst score. Furthermore, it allows the exchange of false negatives for false positives.

Separation

The separation criterion is a criterion of equality of error rates (false positive and false negative). It is like independence but it takes into account the ground truth. A model, in order to respect this criteria, needs to match these two conditions:

- $P(R = 1|Y = 1, A = a) = P(R = 1|Y = 1, A = b)$ (true positive rate)
- $P(R = 1|Y = 0, A = a) = P(R = 1|Y = 0, A = b)$ (false positive rate)

Let's see an example in which separation is not matched. Here the confusion matrix

Gender	R=1 & Y=1	R=0 & Y=1	Tot	R=1 & Y=0	R=0 & Y=0	Tot	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

- P(R=1 | Y=1, A=female) = 3/5 = 0.6
- P(R=1 | Y=1, A=male) = 1/3 = 0.33

As you can notice, the true positive rate is not the same when the gender attribute is changed.

- P(R=1 | Y=0, A=female) = 3/5 = 0.6
- P(R=1 | Y=0, A=male) = 5/7 = 0.71

Also the false positive change is not the same with respect to the variation of sex.
You can see other examples in the slides.

Sufficiency

In the binary case, the sufficiency criterion implies the equality of positive/negative predictive values in all groups. The condition to respect is the following:

$$P(Y = 1|R = r, A = a) = P(Y = 1|R = r, A = b)\ \forall r \in \{0, 1\}$$

Let's see an example.

Gender	Y=1 & R=1	Y=0 & R=1	Tot	Y=1 & R=0	Y=0 & R=0	Tot	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

- P(Y=1 | R=1, A=female) = 3/6 = 0.5
- P(Y=1 | R=1, A=male) = ⅙ = 0.17

Then sufficiency is not respected for positive predictions.

- P(Y=1 | R=0, A=female) = 2/4 = 0.5
- P(Y=1 | R=0, A=male) = 2/4 = 0.5

Here sufficiency is met for negative predictions.

In the COMPAS case only sufficiency was respected, while independence and separation were not.

Algorithmic fairness: some reflections

Reflection 1: fairness and discrimination

We define fairness as absence of discrimination, i.e., the absence of any prejudice, discrimination or favoritism towards an individual or a group based on their characteristics in the context of decision-making.
Note that there are many kinds of characteristics, for example intrinsic vs acquired, absolute vs context-dependent, permanent vs temporary, etc…
Basing on this distinction, we can define different types of discrimination:

- *Direct*: Personal traits of individuals explicitly result in non-favorable outcomes toward them. This kind of discrimination is forbidden by the EU Charter of fundamental rights. Furthermore, there are many secondary laws which go more in detail trying to protect specific groups in specific contexts.

- *Indirect*: Individuals treated based on non-protected attributes; protected groups or individuals still get to be treated unjustly as a result of implicit effects from their protected attributes. This happened in the case in which Amazon decided to offer the one-day delivery service only in some specific zones of certain cities, in that case the ZIP codes were correlated with specific ethnic groups.
- *Systemic*: The discrimination is embedded in the social context, for example an employer which prefer candidates coming from the same region, etc...
- *Statistical*: Decisions are based upon average group statistics to judge an individual belonging to that group. An example is the car insurance case, in which people who drive by night are classified as dangerous drivers because they often go to parties and drive when drunk, but this decision penalizes night workers who use the car mainly for working purposes.
- *Explainable*: Differences in treatment and outcomes amongst different groups can be justified and explained via some other attributes. This kind of inequalities are embedded in the data (example salary with respect to sex). Note that, even if explainable, these kinds of discriminations might be illegal.
- *Not explainable*: Differences in treatment and outcomes amongst different groups cannot be explained and thus are illegal. An example is the Facebook advertisement case.

We can define three characteristics of a discrimination:
- *Nature*: What is the harm and who does it affect?
- *Severity*: For each affected person, how severe or damaging is the harm?
- *Significance*: How many people from a protected group are disadvantaged, and how many in a comparable situation are advantaged?

Usually, we cannot assess the first two, while the last can be measured with algorithmic fairness (note that there are a lot of measures for this aspect, so you need to pay attention to which one you choose).

Reflection 2: No mathematical formalization of fairness exists

There are many formalizations of fairness.
- *Group fairness*: Treat different groups equally (in this course we address only this definition).
- *Individual fairness*: Give similar predictions to similar individuals.
- *Subgroup fairness*: A combination of the two above.

Let us focus for a moment on the individual fairness: any pair of individuals who are similar should receive a similar outcome. But how do we define the concept of similarity between individuals? And how do we state that two outcomes are similar?
Now about group fairness: we can try to reach this characteristic through unawareness, i.e., the protected attributes are not explicitly used in the decision-making process/algorithm. But the Amazon case (in which they use AI to analyze the curriculum of applicants for a specific position) demonstrates that this is not sufficient since a protected attribute could be correlated with another one.

Reflection 3: Mathematical notions of fairness apply only to observational data

You have to keep in mind that all the measures that you use to assess the fairness of an algorithm, are based only on the observed space (i.e., the data you have), which may not reflect the real world.

Reflection 4: Fairness as Justice: equity, equality, need

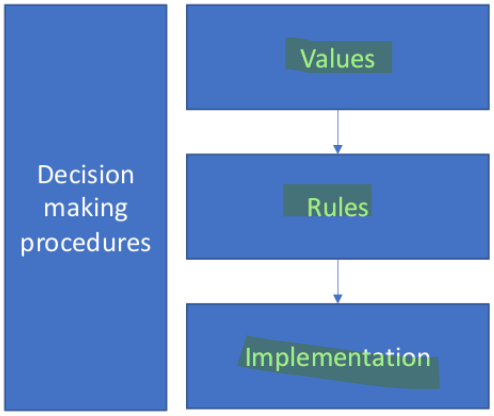
We can interpret “justice” as adherence to law, but also as equality.
When we talk about justice as equality we need to distinguish when there are two private citizens (or legal entities) involved, or a State and some individual person.
In the first case, equality means that the value of the exchanged goods is approximately the same, at the moment of the transaction. Or in a similar sense, if I provoke a damage to someone I have to pay back.

While in the second case, equality is interpreted as equivalence of persons: it means that the State should treat all the people in the same way.
The rule of justice is that principle whereby equals are treated equally and unequal are treated unequally. But how do we define “equally”? In distributive justice there are three main paradigms:

- Equality: assign to each individual the same amount.
- Equity: outcomes are allocated according to contributions/merits.
- Need: outcomes are allocated according to the need of an individual.

Reflection 5: From values to decision making

In the decision making process we can identify three steps, represented in the following picture.



The first is the process in which are decided the values used to create a decision rule, for example which one to consider to grant or deny a loan? The safeguard of the lender’s capital, the opportunity for improving the borrower's life, etc… This decision is taken at the top level management.
The second is the process in which the rules for measuring a certain value are decided. Back to the loan example, if a bank decides to only consider the safeguard of the capital, a decision rule could be a debt/income less than a percentage, the number of references, etc…
While the last layer is the easiest part, in which you take a rule and implement it in the real world.

Note that in exams they ask to explain why a discrimination occurs or if there could be a discrimination and what are the risk factors. They could also ask to reformulate the goal of the algorithm. Let’s see an example: in the COMPAS case, the main problem is that the defendant's point of view was not taken into account, only the one of the judges. So, to reformulate COMPAS, we could remove the computing of the risk score and present to judges only statistics, so that they can infer the score on their own. This is a change in the middle of the decision making process, you cannot classify it in values, rules or implementation.

Fairness qualitative assessment

In this lecture a pattern will be presented to evaluate the impact of an algorithm and to perform a fairness assessment. We can distinguish between two phases of this process:

- Algorithmic risk assessment: assessing possible societal impacts of an algorithmic system before the system is in use, for example the pollution produced by a blockchain or by the training of a deep learning algorithm.
- Algorithmic impact evaluation: assessing possible societal impacts of an algorithmic system on the users or population it affects *after* it is in use.

In order to perform this kind of assessment, you need to answer to these 4 questions:

1. What social conflict arises from the design/operation of a system?
2. What constituencies (interests, values) are adversely affected by this conflict? How?
3. What social interests are at stake in this case?
4. How could the conflict be resolved fairly?

Now we apply this pattern to a real case, then you will have to do the same in the exam with another one.

Case study: safer route

This is a navigator which uses criminal records in order to track the route of the user: if in a zone there have been a lot of crimes, then it will be excluded from the path.

First question

The main issues are:

- Which thresholds should we use to classify a zone as dangerous?
- Were data normalization applied to data? If a zone has a greater density population, then it is normal that more crimes happen
- What kind of criminality are we measuring?
- all the possible bias involved in collected data are embedded into the application (e.g., whether the police monitor more a zone than another)

In the following table, the confusion matrix of this system is reported.

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

Second question

In order to answer this question, we need to identify the stakeholders of the system. You also have to relate them to the previous confusion matrix.

- **Direct stakeholders:** in this case the users of the app, especially those who are new to the city in which the system is used.
- **Indirect stakeholders:** People and businesses living in the areas classified as dangerous, police corps which provided the data for realizing the system.

In case of false positives the zone is classified as “ghetto”, so there is a reputation damage for those who live in that zone. Furthermore, the value of real estates will go down, and shops would have a loss of customers.

On the opposite side, in case of false negatives, users suffer assaults that might otherwise not have happened.

Third question

In this section, you need to identify which social values are involved in the realization of the system. For example, if you are realizing a self-driving system, you have to decide which action to take in an emergency and which people prioritize (moralmachine.net)

Other examples of social values that you maybe need to take into account are:

- *Universal usability*: it refers to making all people successful users of information technologies, for example if you are realizing a web-site you must think about accessibility for those who need a screen reader (blind people), or those who use the interface in a way different than the normal one.
- *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

A more practical example is an online newspaper which stopped its recommendation system, in order to preserve the readers' ability of focusing. In software design, values can be:

- Explicitly supported (in the form of design constraints or formal requirements).
- Brought in by designers, maybe not voluntarily.
- Brought in by stakeholders, i.e., what matters for specific stakeholders.

In the analyzed case of safer route, the social interests at stake are:

- Economic and reputational security: in case of false positives, some areas are wrongly classified as "ghetto"
- Personal safety of individual users: in case of false negatives, some people might be in danger when walking through those dangerous areas classified as safe.

Fourth question

We could solve the issues described before, with some of the following techniques:

- Enable software errors reports and act during operation to calibrate the application.
- Solicit public feedback (also at design phase).
- The need for safety is considered of greatest value and errors minimized accordingly (i.e., we reduce false negatives).
- The economical value of property is considered of greatest value and errors minimized according to it (i.e., we reduce false positives).

Measuring balance in dataset

In this section, we present three measures useful when assessing the risk of discrimination from a machine learning algorithm. Note that these measures are used from the bias audit when performing the evaluation of the system, i.e., they are computed on test data (not labelled). Note that this kind of evaluation is independent with respect to the choices made by designers and developers (e.g., features selection, model selection, …), because the sensitive attributes may be correlated with other proxy attributes (similar to what happened when Amazon decided to analyze candidates' curriculum automatically). Furthermore, be aware that these metrics are only for risk assessment, therefore it could happen that the measured values are good but the algorithm will discriminate some categories anyway (or viceversa).

Heterogeneity - the Gini index

The more equal are the relative frequencies of each category of the attribute considered, the higher is this heterogeneity. It measures how many different types are represented in a dataset. The formula is

$$\mathcal{G} = 1 - \sum_{i=1}^{m} f_i^2$$

where *m* is the number of possible values of that feature, f_i is the relative frequency of value *i*, i.e.,

$$f_i = \frac{n_i}{\sum_{j=1}^{m} n_j}$$

where n_i is the number of times which value *i* appears in the dataset.

With lower values, frequencies are concentrated in a few classes. The minimum possible score is 0, and appears when all the records have the same value. The maximum score is registered when all the relative frequencies are equal, i.e., $f_i = 1/m$. Thus we have that the upper bound is the following.

$$\mathcal{G} = 1 - \sum_{i=1}^{m} \left(\frac{1}{m}\right)^2 = 1 - m \cdot \frac{1}{m^2} = 1 - \frac{1}{m} = \frac{m-1}{m}$$

In order to make this index comparable among different datasets, we apply a min-max normalization as follows.

$$\mathcal{G}_n = \frac{m}{m-1} \cdot \left(1 - \sum_{i=1}^{m} f_i^2\right)$$

Another consideration can be made about those values which do not appear at all in the dataset. When using this index, you can decide whether or not to take into account those values, the important thing is that you stick with your decision for the whole study.

Diversity - the Simpson index

Also in this case, the diversity index is high when values of the analyzed attributes are balanced. The formula is the following.

$$D = \frac{1}{\sum_{i=1}^{m} f_i^2}$$

Here m and f_i have the same meaning as before. This can be seen as the probability of belonging to different classes. The minimum score for this index is 1, and it happens when all samples are concentrated in only one class. While the maximum happens when the distribution is uniform, i.e., $f_i = 1/m$.

$$D = \frac{1}{m \cdot \frac{1}{m^2}} = \frac{1}{\frac{1}{m}} = m$$

Again, we need to make this index comparable for different datasets, therefore we apply the following min-max normalization.

$$D_n = \frac{1}{m-1} \cdot \left(\frac{1}{\sum_{i=1}^{m} f_i^2} - 1\right)$$

Also in this case you can choose whether to consider classes which do not appear in the analyzed dataset.

Imbalance ratio

The imbalance ratio is widely used in machine learning literature. It is the ratio between the highest and the lowest relative frequency.

$$I = \frac{\max_i f_i}{\min_i f_i}$$

Note that with this index you cannot take into account those values which do not appear in the dataset (since the denominator would be zero). This is why in general we do not consider classes with zero frequency. Furthermore, this index is more unstable since if only one class is less frequent it drops.
To make it coherents with previous indexes (i.e., between 0 and 1) we can take its inverse.

SEMINAR CLEARBOX AI

ClearBox AI is a group that offers **synthetic data generation** solutions to help companies preserve data privacy, augment data quality, and automate testing in AI and Analytics projects.

What is synthetic data?
Synthetic data are data that are obtained by generating artificial data that incorporates an original dataset's **statistical properties and distributions**. Clearbox uses the original dataset as a seed to generate synthetic data.

Why should we use Synthetic Data?
Data is the foundation of AI, but there are issues related to:

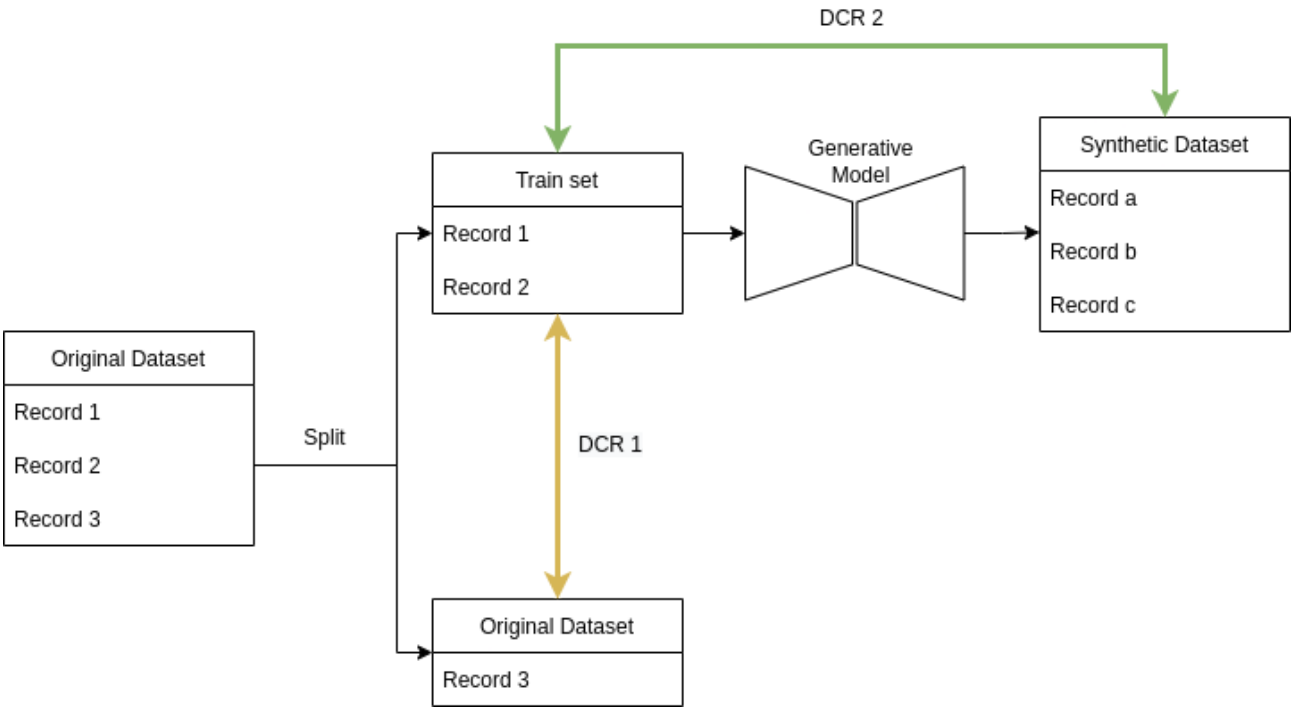
- **Sharing access and privacy of sensitive data** For example in many companies, such as banks, data privacy is so important that it is not possible to share information between departments of the same bank, but at the same time is important to analyze them.
- **Data quality** For example banks have a historical dataset of fraud detection, but they are biased since only the most critical ones are reported, so the data quality is poor.

How can synthetic data be helpful to solve these problems?

- **Protect data and preserve its privacy** Synthetic generation improves de-identification and creates data sandboxes to share data inside and outside your organization easily.
- **Augment data to unlock its most value** Synthetically generated data mitigates data scarcity and improves ML models' generalization by solving imbalance problems.
- **Taking a step forward toward AI fairness** Synthetization is helpful to fix possible bias that lies within the data and ensure a more inclusive AI application.

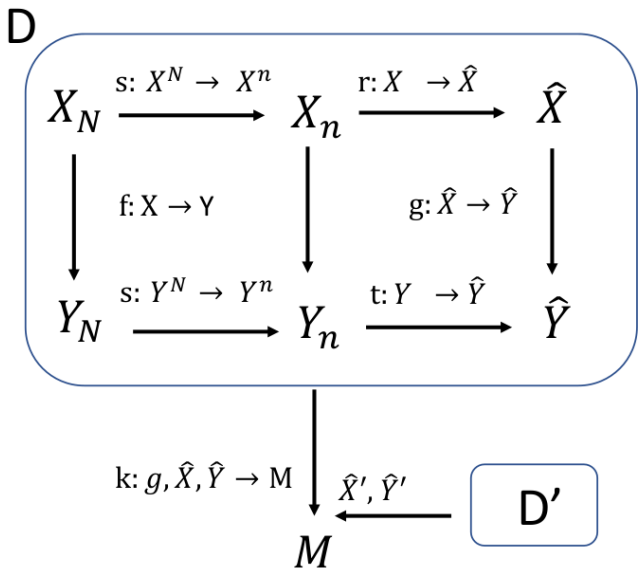
Which measures are used to identify a good synthetic dataset?
A good synthetic dataset does not allow **identification disclosure**, which means an attacker cannot identify an original record using the synthetic dataset. To prevent this problem Clearbox AI uses the **Gower's distance**, which is a way to measure records with different types of data: categorical, numerical, and ordinal. This distance is in [0,1] and it is zero when there is a record in the synthetic dataset that is also in the original dataset.

Given two datasets, the **Distance to Closest Row (DCR)** is the minimal distance between one record of the first dataset and all the points of the second one.
If the DCR between the training set and synthetic set (DCR2) is too close it means that some records could be subject to identification disclosure, while if it is similar to the DCR between the training and test set (DCR1) it means that the synthetic data are interchangeable and so they have generated a good synthetic dataset.



Source of bias

The definition of bias is nowadays an open research topic. In this paragraph, we present some possible sources of bias when building systems which use data. The schema of data is the following.



Here, X_N and Y_N are data which follows the real distribution (impossible to capture in any dataset). Then *f* is the function which we want to learn. Then X_n and Y_n are respectively the features and the target of the samples in our dataset. Then *s* is the function used to do the sampling. Thus, we can only approximate the function which connects X_n to Y_n, not the real one. Then features and labels are projected by functions *r* and *t* respectively: this is made by the data scientist in order to build a good model (which satisfies some measure of success, e.g., accuracy, F1-score, etc…). Finally, *g* is the function learned by the model. K is the training error evaluation function.

During this process, you can have the following sources of bias.

- *Historical bias*: you can have inequalities and disproportion in the real distribution (the one which defines X_N and Y_N). For example, 95% of the best CEOs are men.
- *Representation bias*: this is the bias introduced when sampling data and obtaining the dataset (composed of X_n and Y_n). For example, not all samples are reported, the target variable could be defined in different ways, etc… Some cases are: training an NLP model on samples containing the slang of a specific group of people; training a computer vision algorithm on images taken from specific geographic locations. Furthermore, the behaviour of people change over time, so a sampling done at a specific model may be not representative in another one.
- *Measurement bias*: this is introduced by the data scientist when performing feature selection (subjective choice). Furthermore, granularity and quality of data varies across different groups and may be present proxy variables of sensible attributes.
- *Aggregation bias*: this is introduced by the chosen function *g*, that one used to make predictions. Most of the time, this is caused by the previous source of biases, since it arises when flawed assumptions were made about the population, or there was inconsistency in mapping inputs to labels in the training set.
- *Evaluation bias*: the evaluation data does not reflect the real distribution of the data (this happens because they are sampled too). Furthermore, they may follow a different distribution with respect to the training data, then this influences the score of the model.

ACM code of Ethics and Professional Conduct

Computing professionals' actions change the world. The Code is a document that is designed to inspire and guide the ethical conduct of all computing professionals. The Code includes principles formulated as statements of responsibility, based on the understanding that the public good is always the primary consideration. It is issued by ACM (Association for Computing Machinery).

It is divided in four parts:

1. The first part lists and describes general ethical principles that a computer professional should follow.
2. Section two is about responsibility and it is more practical.
3. The third part is again about responsibility but for those who cover leadership roles.
4. The last section tells about how to be compliant with the code, and what are the implications if you are not compliant.

Preamble

The code addresses every person involved in computing, practitioners firstly (e.g., programmers). It states that they should support the public good in order to act responsibly.

First part

The first article is very important: it states that all computing professionals are obligated to use their skills for the benefit of the society. In other words, you must minimize the negative consequences of the system you are developing, and it should be broadly accessible. The needs of those less advantaged should be given increased attention and priority. Furthermore, computing professionals should promote environmental sustainability.
This part also states that you should be active (take actions) in contrasting discriminations. It states that the use of information and technology may cause new, or enhance existing, inequities.
Another important key point is that you should understand the provenance of the data you use in your project and that you have to think about data minimization (like GDPR).

Second part

You should mantain high standards of professional competence, conduct, and ethical practice. Note that professional competence comes from technical knowledge and with awareness of the social context in which their work may be deployed, and if your application will be deployed in more than one social context, you need to understand all of them.
This part also explicitely mention machine learning systems, which require extraordinary care to identify and mitigate potential risks (since they are based on an inductive process). You should even consider not deploying the system at all (people first, no innovation first).

Third part

This section is designated for formal leaders and informal leader, i.e., people who influence a lot of people with their work. It states that ensuring the public good must be the central concern during your work, regardless of the methodologies and the techniques you use.
Furthermore, as the level of the integration of your system in the society changes, the ethical responsibilities of the organization or group are likely to change as well.

Fourth part

The future of computing depends on both technical and ethical excellence.

Algorithmic accountability act

This is a bill that requires entities to conduct impact assesment on automated decision systems and data protection.

Definitions

1. **Automated Decision System** (From now on ADS): Computational process that makes a decision or facilitates human decision making and impacts consumers.
2. **ADS Impact Assessment**: means an evaluation of the ADS, it includes at minimum:
    - A description of the system, its training data and purpose.
    - An assessment of the costs and benefits of the ADS w.r.t. its purpose, including:
    - data minimization practices
    - period of retention
    - information of the ADS available to consumers
    - extent of the ability to correct or object the results.
3. **Covered entity**: person or corporation on which the Federal Trade Commission has jurisdiction that:
    - Has more than 50M$ average gross revenue in the last three years.
    - Possesses personal information of 1M consumers or 1M consumer devices.
    - Is owned, operated by a person or corporation that meets the previous 2 points.
    - Is a data broker, which means its business is to work with personal information.
4. **High risk ADS**: Is an ADS that
    - poses a significant risk on privacy and security of personal data.
    - poses a significant risk in providing discriminatory decision for customers.
    - analyses or predicts sensitive aspects of life, altering rights or significantly impacting consumers.
    - has a lot of information related to sexual orientation, race, biometric information, criminal conviction, ecc…

5. **Information System**: Is a process that involves personal information, it does not include ADS. It is high risk when involves protected attributes (sexual orientation, race …), monitors a large public physical space.

Data protection authority

The Federal Trade Commision shall regulate the following:

1. Require each covered entity to conduct an ADS impact assessment. For existing systems as frequently as the commission determines necessary. For new systems, prior to implementation.
2. Require each covered entity to conduct a data protection impact assessment of high risk information systems, with conditions similar to the ones of ADS.
3. Require each covered entity to conduct points 1 and 2, if reasonably possible, by consulting an independent external third party.

Automated Employment Tools, Local Law

This is a local law of New Your City regarding automated employment decision tools.

Definitions

1. **Automated employment decision tool**(from now on AED): Any computational process that issues a simplified output to make or assist in employment decisions that affect the natural person.
2. **Bias Audit**: An impartial evaluation by an independent auditor. It should not be limited to testing the tool to assess disparate impact on persons.
3. **Employment decision**: Screen candidates or employment or promotions.

Requirements

1. In the city of new york is unlawful to use an AED for employment decision, unless:
    - Such a tool was the subject of a bias audit within the last year of use.
    - A summary of such an audit is publicly available on the employer website, before the use of the tool.
2. If an employer decides to use such tool on a candidate or employee it must notify:
    - If such a system will be used.
    - Which job qualifications and characteristics the tool will use.
    - If data retention policy is not disclosed on the website, here shall be provided.

Additional Study Material

seminarReport - DIgital Mare Nostrum full report by Francesco Capuano
DATA ETHICS AND PROTECTION FLASHCARDS by Edgar Gaytán