How to answer exam?

Saturday, June 18, 2022 5:20 PM

First Part

Theoretical Question

GDPR Question

1. General description of GDPR
2. Describe pillars appropriate with question
3. Mention what is processing and what is personal data
4. Mention data controller, data subject, ...
5. Mention accountability principle and ...
6. Mention DPIA ...
7. Mention ...

Example

What is the relationship between security requirements and risk management (Article 35) in the GDPR?

(1) The European General Data Protection Regulation in its article 32 describe the security requirements that data controllers must perform when processing personal data in order to comply with the law. Recall that GDPR is a regulation setting a legal basis for the processing of personal data of European Union residents.
(2) The GDPR was developed under a risk-based approach. This means that data controllers must perform an evaluation of the risk posed by the data processing to the data subjects and take proportional and adequate measures for dealing with the risk. Those measures include the security of the data to be processed in terms of technology and procedures used. However, the GDPR does not provide a specific technical requirements because it follows a tech-neutral approach and does not stick into a particular technology that could be disproportionate or not adequate in some cases given also technology development. Instead, it is the data controller that must evaluate the measures to be taken given the context.
(3,4) The data processing involves any operation performed on data linked to an identified or identifiable individual and the data controller must (5) respect the accountability principle. As said before it is the fact it took and adequate measures to minimize the risk by the processing.
First of all a pre-eliminary analysis of the data processing must be done in which the data strategy and management is defined. The risk assessment is done as a by-design approach which means it must be performed before any data processing takes place. Then a formal risk assessment is done, the Data Protection Impact Assessment (DPIA) , that altough is not required in all cases (following the procedural approach) is strongly advised.
In it all the security measures proportionate to the data processing must be described. As suggested by the ENISAA those could be (but not limited to) the training of personal in cybersecurity issues, the physical security of informatic facilities, the procedural design, ...
In short, the data controller must ensure to take secutiry measures appropiate to the risk posed to the data subject. In particular by any unauthorized disclosure, alteration, erasure , .. of the data.
Furthermore the data controller must have (again, if retained necessary) a plan to follow in case of data breach since by the GDPR it is required to notify to the authorities the nature of the data breach and the measures taken.

What is a data breach? What is the procedure to follow in case of a data breach?

A data breach is whenever an unauthorized action such as disclosure,erasure or modification happens on the personal data processed by a data controller.

The concept of data breach and the actions to be taken are discussed in the general data protection regulation of the EU (GDPR). In this regulation a legal framework for the processing of personal data is given. In particular, the responsibilities that a data controller must have in case of a data breach are specified.

Recall that the GDPR is a risk-based regulation, which means that no specific procedure is specified (except for a detail explained in the following about the time of report) and it depends on the context in which it happens. Therefore, in case of a data breach a company should assess the risk of the data breach and inform the Data Authority within 72 hours (this is the only detail specified) giving a report containing the nature of the data breach, its severity and the measures taken to reduce the damages to the data subjects. The data authority is the entity of a member state that ensures that the GDPR is respected.
Recall that the data subjects are the natural persons to which the data is referred (that is data that can be linked to an identified or identifiable data subject) and they must be informed about the data breach only in particular cases where the severity of the breach requires it. Note that this procedure should be done with the help of marketing and legal departments in order to not create excesive alarms and make the company loose money. As an example, we can think of the Unicredit case in which they exagerated a data breach and lost many clients because it was not informed correctly.

The documentation to prepare is extensive and the report to the data authority must be done on time in order to allow it to warn other data controllers about the possibility of an attack (when it is necessary) so a company should have already the documentation prepared in order to not risk to receive a sanction (that can be more or less severe depending on how accountable is the company). The sanction can be up to 20M EUR or 4% of the annual revenue of the company.

The data controller (the entity responsible for the processing of the personal data) must be able to demonstrate that it has taken the adequate security measures to ensure that the risk to the data subject is minimized. As ENISAA suggests they must have taken action in many fronts in order to ensure that there is an adequate level of physical and informatic security. Additionaly, employees should have received training on issues about cybersecurity and good procedures must be adopted to ensure the risk of data breach is minimized.

In the data breach report, the controller must not minimize the data breach that happened. On the contrary it should explain in detail what happened, what are the risks and the measures taken (or to be taken).

Describe the main pillars of the GDPR and how they are related with the DPIA.

The GDPR is the General Data Protection Regulation in the EU. Its aim is to provide a legal framework for personal data processing by enforcing some rules in all EU member states.

SCOPE OF THE GDPR
The GDPR is applicable to the personal data of natural persons in EU soil (so not only EU citizens) (articles 3 and 4). The personal data is defines as any information liked to an identifies or identifiable person. This means that the personal data is referred to a person that could by any reasonable means be identified. It is stated that even if the person could be identifies in the future with technological development it should be taken into account. This is related to one of the pillars of the GDPR that , as we will see, is technology neutrality. The data processing is defined as any moperation such as collection,alteration,storage done on any personal data.
Notice that the territorial principle is not respected by the GDPR in the sense that it is enforcing laws idericetly on companies that can be based on another countries but processs EU natural persons data.

MAIN PILLARS
1) Data-centric regulation : The aim of the regulation is to give the correct importance to the data of the people. So data is in the center of the regulation. It accepts that online identity is important.
2) Procedural regulation: The regulation does not specifically states what are the steps to be taken since it acknwledges that it might depend on the context. This point is strongly related with following points.
3) Risk-based approach : The regulation is based on the accountability principle. That is an entity processing personal data must be responsible for the risk posed to the data subjects by its data processing and the balance of interest must be considered.
4) Tech-neutral apporach : No specific technical measure to be takes is specified since the technology is in constant development and writing a specifi technology could be soon out-dated. Instead the regulation goes to the risk-based approach.

DPIA
The DPIA is related to the risk-based approach. It is a formal assessment of the risk posed to the data subject by the processing of personal data. IN it the company has to clearly state what is the processing done, what are the risks to the data subject and the meaures taken in order to minimize them. The company must also specify the assigned DPO (Data Protection Officer), so the person to verify that the processing is GDPR complient.
The company must be able to demosntrate that it complies with the GDPR principles:
a) Lawfulness of data processing : The processing of the data must be legal and transparent with respect to the data subject. Under this point we find the consent of the data subject that the data controller must be able to demonstrate it obtained following a correct procedure (knowledge of data subject and no asymmetric relationship between data subject and data controller), even if in some cases this is not enough for ensuring the lawfulness of the processing.
b) Data minimisation : The amount of data used must be only the strictly necessary one
c) Security measures : Security measures (Art 32) must be sufficient to compensate for the risk. Adequate technical and procedural measures should be taken
d) Purpose limitation : The data must be processed for the specific purpose for which it was collected.

The DPIA is not mandatory for all applications , however it is strongly suggested. It is mandatory for high-risk applications thus applications involving millions of data or being done by puiblic entities. There is no specific template for it.

AI Question

What are the main issues of regulating AI?

The introduction of new technologies always comes with challenges for the law because problems arise and it must be regulated. However, when we talk about Artificial Intelligence (AI) we observe that it is partocularly problematic to regulate.

First, I will discuss the main problems introduced by the AI. With the introduction of this kind of information systems every piece of information becomes useful data. It can be called the datafication of the society. In fact, every information retireved from every sensor might be used in the context of AI. Take as an example the smart bands that monitor the health of the patient and give a huge amount of data to analyze. This fact itself introduces a challenge since we are monitoring the data and the data is "shaping our society". Another big issue is that before the processing of this data it is not possible to specifically know for which purpose they will be used which is in contrast with the GDPR principles (principle of purpose specification). This is a problem when asking for consent of the users since at the moment of collection of the data it is not known the exact purpose for which it is collected. This can lead to legal issues with current regulations (like GDPR). Nonetheless problems with current regulations are not unique. Current regulations only focus on processing operations but not on individual and group rights (as discrimination, freedom of expression). Key issues of AI is the discrimination of certain groups. This is exacerbated by the fact that those groups are not static , like in the past, but are dynamically defined which does not allow people to associate and fight for their rights. In addition, in most of the cases it is not possible to understand why a certain variable changes the result of the algorithm. The bias can happen because ~~of a~~ bias in the data, but also because of the procedure followed by the data scientist developing the algorithm that could introduce it own belivings into the algorithm through the confirmation bias. Moreover, a AI system can become unreliable in contexts for which it was not trained. Therefore, it is necessary to introduce new laws regulating AI that address those problems.

The first challenge that comes with AI regulation is its definition. In fact, technologies of this field are constantly developing and a narrow definition would lead to companies trying to fall out of the law. It is also challenging to determine who is liable for the "behaviour" of an AI since it is a system that learns from data of users so responsibility is not clear. Secondly, AI is not the same on every field so it is not possible to give an unique law. Possible solutions are:
1) Co-regulation : Hard-law combined with a soft-law in order cover different contexts.
2) Several hard laws for different contexts.

Off course this law must be consistent with existing laws. For example creating a new EU

Discuss how AI should be regulated

Artificial Intelligence has changed the paradigm of data processing. It is technology that processes a lot of data and can be seen as a black box that learns from it and outputs a result or prediction. With the outcome of new technologies new problematics arise and this is the case of the AI. The extensive collection of data necessary for making this systems work contrast with the current regulations on personal data protection already existent that follow the "data minimisation" principle. In addition, there are several aspects not yet regulated about the transparency of those systems and the consequences they might have on society.

As a consequence, there will arise legal issues for which the court will not have the necessary laws to act properly. In fact, the first system that has to face the new technologies are the courts in which the first problems are treated. As an example, think about the HUD case against Facebook in the USA. The HUD is demanding Facebook for the systematic discrimination done by the AI that runs the targeting of the ads in the platform.

DEFINITION OF AI

The first problem we face when regulating AI systems is the definition of them. In fact, it is a technology constantly developing and if we give a narrow definition of it companies would try to fall out of this definition to avoid the regulation. Therefore, a sufficiently wide definition of it must be given.

DATA PROTECTION ISSUES

The AI systems require a huge quantity of information to properly work. However, current GDPR regulations require that the data collected is minimized. In addition, the data subject must explicitly give his consent in order for the data to be processed and he must know the purpose of the data collection. However, at the moment of the collection it is not always known for which purpose the data will be collected with creates a problem that can lead to not GDPR compliance. A possible solution for this would be to use a broad consent (so a consent for many purposes but this would be hard to justify) or a dynamic consent.

TRANSPARENCY ISSUES

Most AI systems act as black box to which certain variables are given as input and an output is returned. Howver, even if we understand which variables have effect in the output it is not always clear why. This can lead to discrimination issues like in the COMPAS case. COMPAS is a system that has been employed in the penitentiary system of the USA to help judges in the decision of conviction of a defendant giving them a score with the probability of recividism. However, the system systematically discriminates black defendants giving them higher scores (thus more probability of recividism). Therefore, companies that develop this systems should perform impact assessments before deploying the system (by-design apporoach) to prevent those issues.

Notice that current regulations only focus on the regulation of how the data is processed and not on group rights (like non -segregation).

For the regulation of this aspect we have to be careful of not regulating against the already existing laws. For example creating another entity that supervise it would not be possible because it would conflict with already existent regulations (GDPR). Instead, an extension of the law should be proposed. In addition, we must consider the fact that not all scenarios are the same and measures to be taken highly depend on the context in which they are applied (the sector). There are two possible routed:
a) Co-regulation : Create a hard law (like GDPR) with several soft-laws (like a code of conduct) for each sector.
b) Several hard-laws for each sector. This option seems hard to implement on practice.

Describe the Jean Monet Chair.

The Jean Monnet Programme is a EU initiative to encourage teaching, research and reflection in the field of Mediterranean studies in computer law. This is done through an interdisciplinary cross-fertilisation contributing to reducing the gap in the field of computer law that affects Mediterranean contries, often underrepresented in the global arena.
The goal is to increase the voice of mediterranean countries in the EUR regulatory policy debate. It will enrich the EU debate by using the significant huminatirian output of the mediterranean countries.

One of the advisors of the JM chair is the representative of homo digitalis.
The association is the result of the fusion between two entities:
a) First one focuses on digital rights, data protection and biometric recognition
b) Second one is about social justice : social pushbacks, non discrimination, arbitrary detention, habear corpus, etc ...

Homo digitalis association has been involved in cases about :
1. Social media analysis for detecting irregular immigration fluxes
2. Monitoring Instant messaging services for detecting migrational fluxes
3. Facial recognition for mass surveilance. Blocked by Italian data protection authority.
4. Behaviour surveillance sustem

In the majority of cases the entities were not doing any DPIA which posed significant threats to the rights.

FIRST INTERVENTION
In the seminar we had during the panel the head of the data protection office of the Red Cross did an intervention.

The Red Cross has as a goal protecting the dignity and life of people which has also a branch in the data protection areas.
Humanitarian activities do not choose the context of operativity as private company operations. Therefore, humanitarian activities have to operate on gihly critical environments. They are regulated by international laws ratjer than national laws because they operate in the best interest of everyone. Usually the international law is a mediation of the national countries laws on this context.

SECOND INTERVENTION

AI can be used to track people crossing the borders. It can certainly improve logistics but the risk is to discriminate against some particular group. Some systems are not mature and inherently biased. The solutions are to:
1. Follow a risk-based approach
2. Respect fundamental right principles
3. Declare which AI applications cannot ever be used
4. Introduce transparency on the AI systems

The intent is to use Ethics with existing regulation. However, with the work of CAHAI , the need for a binding tool with which regulating AI has been strongly needed.

THIRD INTERVENTION
Interaction between humanitarian associations and different private companies who do business in the AI sector.

a) Worls food programme: Palantir provides services of supply chain optimisation powered from AI systems ebedded. The outsourcing of this service is done completely for free for at least 5 years. This makes the world food programme strongly dependent on the WFP.
b) Facebook Libra : A private company serving the role of NGO. For the sake of helping under-developed countries Facebook proposed a crytocurrency system basically leveraging the 1.3B users the platflorm has. It ouwld have resulted in libra threatening global and regionary stability. It would have created a super-power.
c) WHO and Whatsapp : Stopping the widespread of misinformation in the context of covid-19. Whatsapp was the source of communication.

In those 3 cases we see the unbalance between private companies and NGO. NGO need to outsource operations because they do not have the resources. However, this creates a serious risk that the NGO need the private company to operate. The power fo private companies is increased with this partnerships.

ROUNDTABLE

Three speakers gave their opinions about the usage of AI.

The EU border has become the ground of testing for experiments. The migratory flow controls are becoming an actual social control mechanism.
1) The European border control is using avatar for classify immigrants into those that see their aylum request approved or not. Significant risk to human rights because of biases.
2) Emotion recognition application to see areas in which inhabitants are more likely to start riots/protestsallowing pre-control operations

It is expanding EUR power because it is acquiring data from
citizens of other countries. The situation is not humane and it has been normalized. The effectiveness is possible only when human dignity is respected

Other speaker seem confident in the AI usage that allows efficient allocation of financial and human resources.

Discuss current problems of AI trials of regulation

CONTEXT
AI is strongly shaping our society. It could be summarized
society. However, there is no current regulation for the A
aspects of the processing of data (like data protection in
individuals and groups.

PROBLEMS IN REGULATING AI
The main issues when regulating AI are the following:
a) Definition : The AI is a very recent technology that
definition could lead to companies trying to fall ou
compliant with the given definition therefore outsi
b) Transparency : AI systems often act like a black-bo
not clear why. Therefore, the system can systemat
and it is not clear of whom is the liability for it. In f
c) Context : As said before an AI can systematically di
those groups constantly change and are not clearly
cannot defend the rights , because it is dynamically
d) Data protection : The working principle of AI is bas
go against the principle of data minimasation and
current regulations like GDPR. When data is collect
not possible to specify to data subject when the da
which everything can become data for the widspre
about its users

SOLUTIONS
From the problems described above it is clear that curre
cases. Courts have to deal with problems that arise from
correct instruments to give a decision.

With respect to the data protection issues the data cont
usage of data (however, hard to justify in the DPIA) or us
everytime the usage of the data changes.

With respect to other issues a regulation must be added
already existing laws and it is not necessary to write all la
be added. The law must not be the same in all contexts s
innovation so two solutions are proposed:
1. Co-regulation : A hard law (like GDPR) and soft law
2. Several hard laws for different contexts.

CURRENT APPROACH
Currently , there exists a comission of AI experts trying to
several problems with it:
1) The memebers of the commission are not indepen
companies that have a evident conflict of interest.
funded by private companies.
2) Usage of flawless concept: They commission used
3) They create ethical guidance which might be misle
optional when it is required. This is the case of imp
trasnparency (required in another laws).
4) Do not cite any philosphical source which does not
different countries.
5) Ethics and laws must be on different planes:
a. Start from ethical framework and then amar
implement them
b. Reconsider existing laws based on an ethical

The problem with them is that they do not give a clear de
be law-compliant. They give only ethical recommendatio

TOWARDS NEW APPROACH
We should be particularly careful with the following poin
1) AI definition : Now based on the task of the system
2) High risk : A low-risk application can easily become
the text so it does not depend on the context. Curr
technology is considered high risk or not.

It is missing now :
1) Flexibility :
2) Open clause : Defining a close list as high-risk woul
make application fall out of the scope of the regula
3) Decentralized approach : The only comission canno
experts of different sectors.
4) Extend the DPIA

The arise of new technologies make ethical issues arise.

Case Study

1. Actors
2. Material scope
3. Territorial scope
4. Material issues (GDPR)
5. Material issues (AI)
6. Improvement

A Portuguese AI company aims to create a large database for training image recognition algorithms, focusing on objects with different shapes and human bodies. The database will be accessible as a service to third parties on the basis of an annual fee. To achieve this goal and populate the database the company collects worldwide images and videos from blogs, social media, and online media websites. Please consider the characteristics of this case and provide an assessment of potential legal issues and possible related solutions.

(1) In the given case the Portughese AI company acts as a data controller that is processing the personal data of the data subjects, the people that appear in the images that they collect. (2) The images contain personal data about the data subjects in the sense that the body shapes are linked to an identifies or identifiable person. Moreover, body shape is a sensitive data to treat.
The company is based on Europe and it might process personal data of EU residents thus it must comply with the GDPR.

(3)First, we can observe there is a consent issue since the data subjects did not give consent to the data controller to process the data. The data subjects were not informed about the purpose of the collection of the data and if they gave consent to the social networks or websites in which the photos are posted it is unlikely that they accepeted their data to be processed and sell to third parties.

Secondly, we observe that there is a huge risk taken by the company since data associated with millions (potentially) of individual is being processed which in the perspective of accountability principle leads them to having to take adequate security measures. Those measures must be poportionate with the risk taken and in this case it is big since many individuals data is being processed. For example, in case of a data leak the company would be liable for the widespread of the personal data of a huge amount of people.

A possible solution for falling out of GDPR is to anonymize the data so not any personal data is being processed. However, this anonymization procedure promises to be difficult since persons have to be identifies in millions of records and their faces blurried or a similar procedure done. This could lead in bad anonymization and easy re-identification of the persons so it does not seem very promising.

Second Part

Case study

1. Talk about demographic disparities and loop of ML
2. Talk about bias in dataset
3. Talk about bias sources
4. Talk about fairness
5. Discuss ...

A. CASE (7,5points). Coronavirus Relief Package

The Coronavirus Relief Package is a program of the U.S. Department of Health and Human Services that aims at reimbursing hospitals for part of the extra expenses or lost revenues due to the Covid-19 outbreak. Relief funding is allocated according to an algorithm that combines the following parameters:
A) Hospital past revenue (the higher the revenue, the higher the disbursement), measured as:
I) Hospital Operating Margin, i.e. the difference between revenues and costs related to patient care, as a proportion of the revenues (computed on a yearly base);
II) Days of Cash On Hand. i.e. the number of days that an organization can continue to pay its operating expenses, given the amount of cash available (computed at the end of the year);
B) Hospital type (the larger the hospital, the higher the disbursement), measured as:
...
number of daily new hospitalizations higher than 5.000 in the previous year);
C) COVID-19 burden (the higher the burden, the higher the disbursement), measured as:
I) cumulative deaths in hospital due to covid-19 (computed on a yearly base).
II) cumulative hospitalizations due to covid-19 (computed on a yearly base).

The first tranche of reimbursements has been given in April 2021 with data from 2020. A second tranche will be given in April 2022, based on data from 2021. However, the U.S. Department of Health and Human Services declared that the algorithm will be changed, because it has been found that the funding allocation had a disparate impact on Black population. In fact, the audit of the reimbursements given in 2021 at county-level, showed that among counties receiving the same funding, those with majority of Black population had higher COVID-19 burden and worse hospital finances.

Please briefly answer to the following questions (give a separate answer for each question):

1) Provide possible explanations for the results of the audit: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your reasoning (3p.)
2) Which measurement issues do you observe? (3p.)
3) Which data and/or which part of the process would you change to make the impact of the system less harmful in 2022?(1,5p.)

ALWAYS ANSWER SEPARATE QUESTIONS

1) The result of the audit is clear : as a consequence of the lower revenue of hospitals whose patient population is mainly black they will receive less funding because of the algorithm used (higher revenue higher reimbursement).
This is put in light by the audit in which the statistics are shown. Obviously this situation is not fair since we are discriminating a group of individuals because of a protected attribute (ethnicity). The algorithm used to decide the disbursement to give does not use this statistic explicitly (% of black population in a hospital as a variable) , however the hospital operating margin and the days of cash on hand act as a proxy variable for it. A possible explanation of the correlation between these two variables is the already present demographic disparity between black and white people in the US. It is worldwide known that there has been historically a racial segregation against black (and latin) people that have as a consequence lower income thus fragile economical status. Therefore those populations will have cheaper insurances and the hospitals they go to will have lower income thus having lower margin and days of cash on hand. This is further reinforced by the fact that black communities tend to leave together therefore attending particular hospitals and aggravating the segregation.
The algorithm also uses the hospital type and the COVID-19 burden. While for the second variable it does not seem to be any correlation for the first variable it could lead the algorithm to not give enough funds to small hospitals that might be helpful anyway for some small communities. However, this does not seem to be correlated with ethnicity.
2) The measures used, as stated above, act as a proxy variables for the ethnicity of the populations that attend those hospitals. In addition, measures are subjective choices and reflect only a partial reality of the hospital and do not capture important aspects of reality to be taken into consideration. While a hospital could be having more revenue it could be saving less lifes. Furthermore, the measures used close the feedback loop of demographic disparities because they reflect the already segregated reality of society. Then , the model reinforces the demographic disparity because it gives more resources to "white" hospitals further increasing the gap between those two populations
3) A plausible solution for "opening" the feedback loop would be to stop considering economical variables and focusing only on the quantity of people that the hospital is saving. In this way the measure would be more "objective" and hospitals with higher burden would receive more funds. However, we must be careful on not leaving hospitals with small size out of the game since it could introduce a bias against rural populations that have less patients in the hospitals.

B2 (2,5 points). Comment the following sentence: "Technology is shaped by existing power relations".

A. CASE (7,5points). Comment on the J-Crime Prevention case.

The tool J-Crime Prevention (J is for Juvenile) provides judges with a risk indicator on young defendants (i.e. less than 18 years old). The tool is consulted by the judges who will decide which sentence to apply (prison, parole, social work, etc). The indicator estimates the probability that the person will commit a new crime within the next three years, and it labels him/her as at risk of recidivism or not. The tool has been adopted by a certain number of courts in Italy.

An independent audit has been carried out to check whether systematic discriminations occur. The researchers who made the audit used a metric called demographic disparity (DD), which is computed as PP_i / PP_r where PP is probability of being predicted positive for recidivism, for members of group i and r. PP is computed as the number of defendants of the same group predicted positive for recidivism divided by the number of all defendants of that group. For example, $DD_i = 1$ means that someone with attribute a_i is just as likely to be classified as recidivist as someone with attribute a_r . $DD_i = 2$ means that someone with attribute a_i is twice as likely to be classified as recidivist as someone with attribute a_r . The results of the audit are:

- DD_{female} = 0.75 (reference group: male)
- $DD_{foreigner}$ = 1.90 (reference group: Italian)

The company that developed J-Crime Prevention declared that its algorithm does not use neither gender nor nationality in its algorithm, but only the following information:

a) *Previous violent offences*: the number of violent offences previously committed
b) *Previous non-violent offences*: the number of non-violent offences previously committed
c) *Criminal parent/caregiver*: whether who took care of the defendant in his/her childhood had committed a crime
d) *Poor school achievement*: whether performance at school has been poor in at least 3 years
e) *Poor compliance*: whether defendants lived most of his/her life in poverty

Neither the source code nor the documentation of the algorithm is public, hence it is not possible to know how the variables were used (e.g., their weights). However, after the audit results were published, the company published the recidivism rates in the training data: 45% Foreigners vs. 30% Italians, 40% Male vs 20% Female.

Please provide possible explanations on why J-Crime Prevention produced rankings that disadvantaged male defendants and foreigners. Clearly state your own hypotheses, and any other information that you suppose, in order to coherently support your reasoning.

We observe that the J-crime prevention system developed by the company presents a bias against the foreign and male defendants. This bias is expressed by the fairness measures used by the independent audit company in which we observe that foreign defendants have almost twice of likelihood to be labeled by the system as re-offendants and the females are also advantaged.

A first explanation of the result of the audit is the fact that the company admitted to have used for the training of the algorithm a dataset that has a bias. In fact we can see that foreign and male persons have higher percent of recidivism. This bias can be due by a demographic disparity that is propagated to the data. In fact the reality in italy is that usually the foreign population have a disadvantaged economical situation with respect to italians which can lead to higher criminality rates.

Secondly, we can see that even if the algorithm does not use nationality as a variable it uses a range of variables that are correlated with this factor thus can be considered proxy variables. As previously stated the poor compliance and poor school achievement variable can be strongly linked with the nationality of the defendant.

Furthermore, the correctness of using variables as criminal parent could be argued to be not ethical since we are taking an individual decision based on factors not directly controllable by the indivdual
The most plausible hypothesis that links men with reicidivism is the fact that most of people that commit crimes are men because of historical inequalities for which men have to get the money for the house. It can also be due to a measurament bias, this is the fact that violent crimes (that we suppose are committed prevalently by men) are better reported than not violent crimes (more committed by women as hypothesis) which leads to an imbalance in the dataset thus to a discrimination done by the classifier.

The risk of this classifier is to further increase the demographic disparity between those groups by reinforcing the underlying disparity that is already present in our society and aggravated by the measures taken in order to train our algorithm. This algorithm if deployed would close the feedback loop and increase the disparity.

B1 (3,5 points). The following contingency table refers to results of a classifier that outputs:
- R=1 when a post on Facebook is deemed offensive and is obscured
- R=0 when a post on Facebook is not deemed offensive

The target variable is Y and it indicates when the post was truly offensive (Y=1) or not (Y=0).

Option 1

a. Comment the data in this particular context (hate speech in social networks) (1,5 points)
b. After a change in the algorithm, the table is updated: 40 more cases are added to the cell {Female, R=1 | Y=1} and 40 more cases are added to the cell {Male, R=0 | Y=0}. What will change in terms of reported fairness? Why? (2 points)

	Y=1			Y=0			
Gender	R=1	R=0	tot	R=1	R=0	tot	TOT
Female	20	40	60	20	40	60	120
Male	60	40	100	60	40	100	200
tot	80	80	160	80	80	160	320

Our classifier wants to predict wether a post is offensive or not. By the data collected on the performance of the classifier we can compute the fairness criteria in order to state if the classifier is biased against by certain group of population by a protected attribute as it is the gender.

I will focus on using the separation criteria that states that given a value for the protected attribute and a value for the predicted variable our algorithm should provide predictions with the same probability.
P(R=1|Y=1 and A=male) == P(R=1|Y=1 and A=female) (the true positive rate must be the same)
We have P(R=1|Y=1 and A=male) = P(R=1 and Y=1 and A=male) / P(Y=1 and A= male) = 60/100=0.6 and P(R=1|Y=1 and A=female) = 20/60=1/3=0.33

As we can observe from the data computed in this way we see a disvantage for men. It is also possible to see the false negative rates for women.
FP_male = 0.4 FP_women=0.66

We can further comment the results by saying that the dataset used for testing is perfectly balanced.

b) In the described case the true positive and the false negative rates for both classes are perfectly balanced thus the fairness test is respected by looking this part of data.
In addition we see almost a balance in terms of the false positive given to both classes. While it is still higher for men the difference is only 2/21 between the two classes.

1)
Context
The PESO software described is a software that will have a strong
life of the work seekers of Spain and will define how the public r
helping them will be distributed. The software aims to choose c
to increase the efficiency of the job consultancy services in Spain
effectiveness. It is of crucial concern that this software will not s
discriminate against certain groups.

Measurement Issues
The variables used are very problematic and could potentially lea
discrimination against workers that cannot easily find jobs even i
system is to help them to be introduced into the Spanish labour

a) Gender : The gender is a protected attribute of many legisl
the European one and no person should be discriminated l
does not seem any reason why a person should receive dif
on the gender. This variable raises not only a ethic concer
one.
b) Age group : Persons from specific age groups could need s
finding a job. To give an example old generations are speci
labour market because (for example) their knowledge mig
date or could have lower performance than younger work
lead to a systematic discrimination against elder people. A
fragile group are the youngest one that because of the no-
have trouble in finding a job and be classified as in A2 clas
difficult labour market insertion. This kind of systematic di
would be against the public good.
c) Occupational group : Altough this might be a valid predicto
treated with extreme caution to not over-help one of the t
However, it is very likely it would help to solve inequalities
working in those two sectors.
d) Assistance obtained in previous years : This variable could
harmful because it would deny help to people that need it
that received help in the past years could belong to profes
labour market insertion might be particularly difficult.

2)
A better goal would be to measure effectiveness of the program
amount of people in difficult situation helped and identify those
based on data, require more help for finding a job. For defining t
situation it is enough to keep as measure the age group, the assi
and occupational group. Definitely the gender feature should be

Theory

COMPAS case

0
The COMPAS case is a clear example of the negative effects an algorithm can have in a society and how it can "close" the feedback loop of demographic disparity. It is an algorithm that the US penitentiary system started to use to help the judges in taking decisions about the detention of the defendant prior to the trial or when sentencing. It gives defendants a score from 1-10 that defined the probability of it to re-offend the law based on 137 characteristics, among which ethnicity is not present. What independent audits of the software found is that the algorithm is much more likely to give a bad score to black defendants with respect to white

d in the following sentence : software is shaping
AI. Current regulations focus only on procedural
GDPR) and not on fundamental rights of

has arised in the recent years. A very narrow
t of the regulation by producing systems not
ide of the scope of the regulation.
x. Certain variables influence the result but it is
ically discriminate against certain populations
act, it is a technology that learns from users.
scriminate against certain groups. In addition,
defined so it is a new concept of group that
defined so people cannot group to fight.
ed on the usage of a lot of data. However, this
purpose of collection fo the data expressed in
ted it is not clear what will be the usage so it is
ata is collected. In addition, we live in a period in
ead of "smart" devices that collect information

nt regulations are not enough to cover all this
the usage of this technology and should have

rolles might request for a wide-consent for the
e dynamic consent that requesta authorization

. However, it is necessary to be coherent with
aws from zero. A new piece of regulation must
ince this would be too restrictive and stop

s for different contexts (like a code of conduct).

o regulate it in the EU. However there are

dent. They come from big technological
In addition, academic researches are also

the concept of "trusting AI" which is irrational.
ading because companies might think that it is
act assessments (required by DPIA) and

take into account the different context in

ate laws expressing how companies should

frameowork.
ocument helpful for AI companies that want to
ns.

ts :
n. However, it evolves rapidly.
high-risk. The concept of high-risk is defined on
rently there is no assessment to explain why a

d result in companies trying to strech the law to
ation
ot emanate the regulation without thte help of

Describe again the problems in regulating the AI. How is it related with the seminar given in class.

The AI is a new technology from which new ethical and legal issues arise. First we consider the issues with respect to the already existing regulations in data protection. Secondly, we will consider the ethical issues deriving from the nature of the technology.

DEFINITION ISSUES
It is necessary to give a correct definition of the AI in order to correctly regulate it.This is challenging because it is a new technology constantly developing in a very fast way. The risk of giving a too narrow definition of it is companies stretching the law in order to fall out of the scope of the regulation. It is a problem already present in the EU commission designated for regulating the AI. In fact, they define a high-risk AI with a closed clause which can lead to this kind of problems.
DATA PROTECTION ISSUES
The current regulation in data protection in the EU is the GDPR. It requires to respect the principles of data minimization and specification of purpose for the processing of personal data. In contrast, the AI technologies often require to use millions of records to correctly work or to obtain good performance. Furthermore, it is not always known the purpose of the data collection at the moment in which it is effectively collected.
Possible solutions to this second problem are to request from the user a wide-consent or a dynamic consent that changes as the purpose of the data usage changes in order to be compliant with the law.

TRANSPARENCY ISSUES
The AI are systems often considered as black boxes, in the sense that a varibale is known to change the result of the algorithm but the reason is not clear. The problematic is increased by the fact that the systems can be biased and systematically discriminate against a certain populationor group of people.
A new problem arise when considerin AI. The groups defined by the AI dynamically change which makes virtually impossible for people falling in disadvantaged groups to "group" together and fight for their rights, as done historically whenever any systematic discrimination was done.
In addition, it is not clear who is liable for the behavior of a certain algorithm given that AI learns from the users.
Therefore, it is necessary to regulate it fro giving a clear guidance to companies using this systems on how to assess the risk posed by them to individual focusing on individual and group rights and not only on the processing rights as done by current regulation like GDPR.

NEED FOR A LEGISLATION
By the arguments given in the previous paragraphs it is necessary the regulation of the AI. However, it is not possible to overwrite existing laws. Instead, it is necessary to extend already existing laws for not creating legal paradoxes. For example, it was purposed to create a new organism in the EU for the regulation but it could interfere with the data protection organisms so it is not possible.

Secondly, the creation of a single law for regulating all AI systems completely ignores reallity in the sense that depending on the context of usage a certain technology might be less or more dangerous and there are no absoulted. Therefore the best solution for the AI regulation is :

a) Co-regulation : A single hard law (like GDPR) with several soft-laws (like a code of conduct) depending on the sector
b) Multiple hard laws depending on the sector

NEED FOR MEDITERRANEAN PARTICIPATION

As we see that AI regulation is necessary to understand that the devolping of this laws must be done by a commission composed by several experts from different fields. Currently, the EU commission for AI regulation is composed by people coming from big technology companies and the few academics present are funded by private companies so the risk of bias is high. In addition, current proposed regulation lacks of a philosophical basis and does not take into account the context of different countries.
In this context, we should consider that the Mediterranean countries have a high humanistic output and have been largely underrepresented on the global arena in the field of technologies. The Jean-Monet Chair has exactly the purpose of incresing the participation of Mediterranean countries and make their voice stronger.
Members of its chair range from people working in the Red Cross and in Data Rights Protection Associations.

Just to give an example we can think of one of the members of the chait that comes from Homo Digitalis. This is an association composed by the fusion of a social justice association part and a data protection part. They have worked in the context of usage of AI for preventing irregular immigrational fluxes by monitoring instage message applications and social networks.

In the seminar held in our lessons, during the intervention we saw how the private technologies companies try to associate with humanitarian organizations in order to increase their power. For example, Whatsapp tried to associate with the WHO during the covid-19 pandemic for stopping the spread of misinformation. Ultimately, this would have increased the power of Facebook.

and explain how to interpret the result (1,5p.)

Class	N_i
African-american	16
Asian	12
Caucasian	22
Hispanic	4
Other	12

Class	N_i
African-american	16
Asian	12
Caucasian	22
Hispanic	14
Other	12

Class	N_i
African-american	6
Asian	12
Caucasian	22
Hispanic	4
Other	12

Class	N_i
African-american	16
Asian	12
Caucasian	22
Hispanic	40
Other	12

Class	N_i
African-american	16
Asian	12
Caucasian	22
Hispanic	4
Other	12

(uguale a Tab.1)

Class	N_i
African-american	16
Asian	12
Caucasian	12
Hispanic	4
Other	12

The gini-simson index is a measure of balance. This means it is helpful to measure how heterogeneous a dataset is. In particular it is described by the formulas :
G=1-sum(fi^2) where fi is the relative frequency of the class I and we sum up over all classes
Gn=(m-1)*G/m
This is the normalized gini index : it is normalized by the number of classes present in the dataset

The value of the gini index reflects how balanced is the dataset. In particular, the higher it is the more balanced is our dataset considered. A perfectly imbalanced dataset has a gini index equal to 0.

G = ..
Gn = ...

As we can see from the result our data is not perfectly balanced. This can be explained by the fact that hispanic people has a low relative frequency with respect to other categories like caucasian. Therefore, they could be underrepresented. This might lead to bias problems if used in the context of ML.

g impact on the
esources for
andidates in order
and increase their
ystematically

ad to a systematic
f the aim of the
market.

ations, including
based on it. There
ferent help based
but also a legal

pecial help in
ally fragile in the
ht be not up to
ers. This could
nother specially
experience could
s thus leading to
scrimination

r it should be
wo groups.
between people

potentially be
. In fact, people
sions for which

based on the
professions that,
he difficult
stance received
removed since it

A. CASE (7,5points). Prioritizing patients for intensive care programs

The health system of USA relies on a risk-prediction tool to target patients for "high-risk care management" programs, where identified patients are treated with additional resources. These programs are widely considered effective at improving successful treatment outcomes and satisfaction while reducing costs. Because the programs are very expensive (dedicated nurses, extra appointment slots, etc.), the USA health ministry uses the risk-prediction tool to identify patients who will benefit the most. To achieve such a goal, two key assumptions are made: i) those with the greatest care needs will benefit the most from the program; ii) those with the highest medical expenditures are those with the greatest care needs. Under these assumptions, the targeting problem becomes a pure prediction problem: predict future medical expenditures of patients.

Developers built the algorithm relying on past data to build a predictor E_{it} : total medical expenditures for patient *i* in year *t*. The algorithm is given a dataset with the following data on care utilized and billed to patients *i* over the year *t-1*

a. Demographics: age, sex.
b. Health insurance type: basic insurance provided by USA or commercial insurance.
c. Code of the main illness diagnosed.
d. Prescribed medications.
e. Appointments made, with type of medical service (e.g., surgical, radiology, etc.).
f. Billed amounts, categorized by type (e.g., outpatient specialists, dialysis, etc.).

It is not known how these features were combined to make the prediction. An independent audit looked at the algorithm's results (patients identified for the intensive care program) from the point of view of race. The published analysis results are reported in the following table:

	White	Black
Nr of patients	45000	5000
Average age (years)	55,5	45,5
Female (%)	60%	70%
Race composition of the intensive care program (%)	80%	20%

1. Provide possible explanations for the results of the audit: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your reasoning. (3p)

2. Which data and/or which part of the process would you change to better achieve the goal of the system? (2,5p)

1) The ultimate goal of the algorithm is to provide people with the greatest care needs with a program that will provide them with additional resources. However, as shown by the audit the algorithm presents a huge bias against dark skin color people. In fact almost 80% of the receipients of the program are white while only 20% of them are black. Notice that the algorithm does not use ethnicity as an explcit variable.
The bias can be attributes to the fact that an assumption is done about the people that need the program : they should have great helth expenditure. keeping aside wether this assumption is ethical or not it acts, withou doubt, as a proxy variable for the ethnicity.
US is a strongly divided society in whcich a historical demographic disparity against dark skinned people is present. This group of people are socially diasvantaged and have less economical resources than white people (generally). In addition, this group of people has been historically discriminated explaining that it can use less the health resources. Among the predictors there are several proxy variables for the economical status thus the ethnicity which leads the algorithm to be biased. I will discuss each variable in the following:

Proxy variables:
a) Demographics : The algorithm uses the age and the sex of the person. While the age can be a determining factor in deciding if a person need special healthcare (think about elder people) special attention must be taken to not exclude potentially toung people with serious illnesses that also would benefit from the program. The gender seems to not be relevant and , as shown in the data, unfairly influences the predictions
b) Health insurance type : This is a clear predictor correlated with the economical status of a person. It acts as one of the above descibed proxy variables.
c) Code of main illness diagnosed : Important factor, does not determine discrimination.
d) Prescribed medication : Important factor, does not determine discrimination
e) Appointemnts made : While it can be determining for understanding the severity of the illness it can act as proxy variable since white people can easily go to doctor because of the better insurance and the freedom of time.
f) Billed amounts : Clear proxy variable that discriminates against not wealthy people.

2) As said in the previous paragraph the model is based on a questionable assumptiuon that the people that need the most additional healthcare resources are the ones that have spent more money on it. Unfortunately, the world has the demographic disparity descreibed above so this assumption is trivially false. A possible change would be to eliminate this assumption and try to give the healthcare to the people that need it the most. In fact, people that would benefit the most are not wealthy. All variables correlated with financial status should be deleted, and where this is not possible (like with the appointments with specialist made) compensate the fact that this could lead to the increase of the disparity.

A. CASE (7,5points). Student Success Predictive Model.

The company ABCsoftware.com (in short: ABC) sells a service based on the Student Success Predictive Model (SSPM) to predict the graduation likelihood of university students in the USA. Once the software produces predictions, no further guidance is given to professors who serve as student advisers, and it is up to the school to decide what to do with predictions: typically, the service is used by universities to recommend students changing major or to withdraw their career.

Predictions are made with a university-specific model: a customized set of predictors is constructed from student records of the specific university, and then combined and weighted using an automated training process designed to maximize predictive accuracy. Historical student records that are used in the training set satisfy the following criteria: i) matriculated between 1st Sept. 2005 to 31st July 2010; ii) had at least three registered successful exams passed; iii) yearly subscription taxes regularly paid. Predictions are made on any type of student, i.e. from incoming freshmen to nearly-graduating seniors, and a final rank is provided with success scores from low to high.

Among the predictors used for training the model in University X, the ones which had higher impact on the score of the software (i.e., responsible for more than 5% of the variance in scores) are the following ones:

A. *Admit Code*: a student's admission type, which can be: first time freshman, transfer from another university, conditional admit (documentation missing)
B. *First Generation Indicator*: "Yes" or "No" indicator of whether any of an individual's parents have ever earned a bachelor's degree.
C. *Median Income by Admission Zip Code*: The median household income in the zip code of a student's home at the time of her admission
D. *High School GPA*: a student's high school grade point average
E. *Legacy Indicator*: "Yes" or "No" indicator of whether a student is someone whose parent or other family member attended the same college.
F. *International Indicator*: Yes" or "No" indicator of whether an individual is an international student
G. *High School Percentile*: a student's high school rank in terms of percentile (in that specific high school)
H. *High School Size*: the size of an individual's high school student body
I. *Veteran Indicator*: "Yes" or "No" indicator of whether a student is a veteran of the United States Armed Forces
J. *In State Resident Indicator*: a "Yes" or "No" indicator of whether a student is a resident of the university home state.

An audit found large disparities in the software outcomes for students of different races, according to the following table:

Table 1

	Black	White	Latino	Asiatic	TOT
Success	2500	6500	1200	2400	12600
Moderate success	1050	3500	650	2100	7300
Failure	1300	220	1100	260	2880
Unknown	10	50	30	30	120
TOT	4860	10270	2980	4790	22900

Table 2

	Black	White	Latino	Asiatic	TOT
Success	2200	6600	1100	2300	12200
Moderate success	1350	3400	650	2000	7400
Failure	1600	210	1100	260	3170
Unknown	10	45	30	25	110
TOT	5160	10255	2880	4585	22880

Please briefly answer to the following questions:

1. Provide possible explanations for the results of the audit: clearly state your own hypotheses, and any other information that you suppose in addition to the provided data, in order to coherently support your reasoning. (3p)

2. Which measurement issues do you observe? (3p)

3. Which data and/or which part of the process would you change to make the impact of the system less harmful? (1,5p)

The ABC system is used for giving a sensitive suggestion to students about their career. The Audit found results that show that based on the ethnicity of the student the software gives different success scores. For measuring it we can use the fairness measure of independence and compute for each ethnicity the probability of it being classified as a success student given the ethnicity.
P(success|white)=0.63
P(success|black)=0.51
P(success|asian)=0.51
P(success|latino)=0.40

As we can observe from the computation based on the data of the audit the systems tends to give and advantage of score tho white students (10% difference) and gives lower scores to the latino ones. However, the system does not use ethnicity as a factor so the variables that are used must be correlated thus act as proxy variables for this protected attribute (in many legislations it is illegal to discriminate based on ethncity).

VARIABLES

As a context we must state as hypothesis that economical status is linked to the ethnicity. In fact, in the US there is a historical demographic disparity in terms of economical status between white and other ethnicities. Hisotrically black and latinos have been disadvataged and population of immigrants that, unfortuanately, do not enjoy of the same economical status as white people.

Admit code : The type of admission code. This factor does not seem to be linked with ethnicity.
First Generation Indicator : This indicator is higly linked with the economical status of the family. In fact, richest families will have parents with at least a Bachelor Degree. This factor is linked then with the ethnicity.

1. Please briefly comment the results of the audit . (5p)

2. Which other ethical issues do you observe, including measurement issues? (2,5p)

1)
CONTEXT
The CheckTax software used in France will have strong impact on the society because it will give the state a tool for classifying citizens as potential tax evaders. This will lead to further examination. For ethical reasons this process should be done with no ethnical discrimination. Therefore, the audit is necessary as an impact assessment before deplyment of the software. The fact it is tested on past data an not "on the fly" is a good practice since it is not having an impact on persons.

RESULTS
The audit shows that the algorithm is biased against French citizens. In fact we can compute one of the fairness metrics in order to check if the protected attribute has any effect on the result. I will use the separation criteria: the true positive rates and true negative rates (or false positive equivalently) should be approximately the same.
TP : When considering the true positive rate we have for french citizens 600/800=0.75 while for foreign citizens we have 100/1000=0.1. There is a huge gap between this metrics showing that the algorithm incorrectly classifies many foreign citizens as not tax evaders while is very strong at "judging" french citizens.
FP : Repeating the computation we have that for french citizens it is 8/26 (false positive) while for foreigns the metric is 7/25. In this case the false positive rate is roughly the same.

The huge gap in the results for the true positive rates can be explained by the working principle of the algorithm.

First, we have to consider the fact that for determining the not consistent economical status of a person we are considering the behaviour on social networks and e-commerce portals. However, this is based on the assumption that both populations have the same behavior and spend the extra money in the same way. Those two ethnic groups have different behaviours and in particular will probably spend the money in different websites which could lead to a disadvantage of French people that use the French websites with which the government can have an agreement.

2)
The algorithm does not seem to preserve good the right of the citizens to have a privacy. While it could be argued that it is a proportional measure taken by the state to avoid tax evasion so it respects the balance principle it is not clear if the users gave the consent for their data to be processed in this way by the websites. Therefore, this measure could lead to a legal issue (GDPR) and also arise ethical issues about how far can the state try to "observe" its citizens. Additionally, as said before , there is the flawless assumption about the same online behavior of different ethnic groups which does not consider diversity of behaviors and penalizes people that uses this portals.
By Campbell's law the observation that the usage of those websites leads to further investigation of the personal finances by the state could lead people to simply stop using those websites and use other online e-commerce sites that are not affiliated with French government. Finally, this can lead to e-

A. CASE (7,5poi
CheckTax softwa
data on citizens c
agreement with th
economic well-be
Depending on the

R=1 if the person
R=0 in the opposi

The experimentat
combining it with

An algorithm aud
nationality and ci
where Y=1 indica
Table 1

French
Foreigners

An Italian ca
repaying the
to the applica
2010 to 2020
form with da
that data, app
used by the c
In fact, natio
and they also
following po

The features

A. Birth
B. Age
C. City
D. Car:
large
E. Clai
F. Year

The analysis
significantly
Therefore, th
respect to tra

Please briefl
QUESTION

1. Provide po
and any othe
reasoning. (3

2. Which me

3. Which cha
terms of sepa

After the te
based on th
the text, w
there must

Before ana
some histo
groups (bla
addition, w
them even
Italian citiz
No informa

nts), Software CheckTax.

re is being tested by France to classify citizens as potential tax evaders or not. It works with collected both from social networks and from various e-commerce sites that signed an e French State (e.g. railways, airlines, amazon, etc.). From this data, an indicator of ring is developed, which is then compared to the average of the last three tax returns. result of this comparison, the software returns:

is considered a potential tax evader
te case

ion is done in the department of Alpes-Maritimes using data between 2014 and 2017 and data on tax evasion controls, available from the Ministry of Economy and Finance.

it is made to verify that no systematic discrimination occurs between citizens of French tizens of other nationalities (but still residing in France). The results are shown in the table, ates a person on whom there is at least one verified tax evasion, and Y=0 otherwise.

Y=1		Y=0	
R=1	R=0	R=1	R=0
600	200	8000	18000
100	900	1400	3600

r insurance provider experimented the usage of machine learning for predicting the risk of costs of caused incidents in a year for a person. In case of high risk, a higher premium is offered ut (who can accept the quote or not). The company used part of its own historical data (from) to train a classifier. During the experimentation, applicants could get their quote after filling a n about the car and themselves (fields filled are the same used by the classifier). In addition to plicants were required to input information on ethnic group: although this sensitive data was not lassification algorithm, it was collected in order to test the classifications against discrimination. al and international regulations require that prices should not vary depending on ethnic group forbid training of classification/prediction algorithms with it. The ethnic group had the ssible values: Caucasian, Black, Asiatic.

used by the classifier are the following ones.

place: driver's nation of birth (list of all possible countries in the world)
driver's age (integer between 18 and 100)
driver's residence (list of all Italian cities)
insured vehicle type, deduced from the car model and year (two possible values: small cars and powerful cars)
n history: number of previous claims (values: less than 3, between 3 and 6, more than 6)
y distance: estimation of kilometers driven in a year.

of the data gathered during the 6-months experimentation showed that offered prices varied when only ethnic group differed while all other characteristics were equal.
e company dismissed the development of the system, and kept applying insurance costs with ditional, static, risk models.

answer to the following questions (GIVE A SEPARATE ANSWER FOR EACH
:

ssible explanations for the results of the experimentation: clearly state your own hypotheses, r information that you suppose in addition to the provided data, to coherently support your p)

asurement issues do you observe? (3p)

nges in the experimentation data collection process would you introduce to check fairness in ration? (1.5p)

sting of the application we can conclude that the machine learning algorithm has dome bias he ethnic group of the person requesting for insurance. However, this variable, as stated in as not used for training the classifier. Therefore, we conclude among the used predictors be at least one variable with a correlation with the ethnic group.

lyzing in detail the variables we can understand the context of the Italian society. There are rical demographic disparities affecting mostly immigrant population. Usually those ethnic ck and asiatic) have less economical resources than nationals (usually white people).In e could hypothize that people that have been living less time in Italy have less claims against if they had an accident in other country. This situation could lead to the penalization to the ans (mostly caucasian).
ation is available about the training set. A balance test must be done in it (and this hypothesis ated)

0

The COMPAS case is a clear example of the negative effects an algorithm can have in a society and how it can "close" the feedback loop of demographic disparity. It is an algorithm that the US penitentiary system started to use to help the judges in taking decisions about the detention of the defendant prior to the trial or when sentencing. It gives defendants a score from 1-10 that defined the probability of it to re-offend the law based on 137 characteristics, among which ethnicity is not present. What independent audits of the software found is that the algorithm is much more likely to give a bad score to black defendants with respect to white defendants showing a clear bias against this population thus systematically discriminating them.

It is a black box system in the sense that the algorithm and datasets are not available publicly. However, its behavior can be explained based on the demographic disparity present in the US that introduces a historical bias to the algorithm. US is a highly divided society and for decades afro-american population has been discriminated and has been characterized by a series of economical and under representation issues. Therefore, historically, afro-american families have lower economical status.

This algorithm closes the loop in the sense that gives ,again, an advantage to white defendants even if , when analyzing the fairness criteria, it gives much more false negatives when analyzing white population than with black population. The model, influenced by a series of historical biases and development biases, disadvantages a historically discriminated population.

If we want to solve this problem so that both populations get the same false negative rates (or false positives, depending on the criteria used) we would have to use different thresholds. However, by US law, this is not possible since ethnicity is a protected attribute and no discrimination can be done by it. Therefore, we have a problem with no solution.

If we stick to the ACM code we can conclude that a system like this should not be deployed because of the risk that it poses to society and it is not on the public good interest. This case also highlights the necessity to regulate AI systems that can shape the society and the company is not liable for a potentially harmful system. The company should be accountable for it and the software should be audited before deployment in such a way that the risks posed by it could be evaluated properly and the unfair biases introduced can be properly addressed.

However, we must be careful on not leaving hospitals with small size out of the game since it could introduce a bias against rural populations that have less patients in the hospitals.

B2 (2,5 points). Comment the following sentence: "Technology is shaped by existing power relations".

B3 (1,5 points). What are main allegations to Facebook for its advertising platform?

B2) Our society carries a historial demographic disparity. In fact we can see how the people in the power of the society have stayed in that position for a long time and have shaped the society to their advantage. This is perfectly described by the concept of the matrix of oppression. In fact, most powerful people are mostly of white ethnicity and male gender leaving all other categories historically forgotten and opressed.

As an example on how those demographic disparities we can think on the facial recognition systems that , as shown by several studies, perform better for light skin people. This is because people with higher economical resources have this type of skin leading to an intrinsic discrimination of black people. Datasets collected and on which algorithms are tested are for white (men) people rather than black and women. A confirmation bias is introduced by the fact that people working on those projects have higher econmical means thus more likely belong to those ethnicities (people that go to university and educate as software engineers/computer scientists usually do not come from lower part of society). As we can see the existing power relations shaped a new technology.
Another example we can give is the fact that hiring systems tend to prefer male workers, univeristy entrance systems prefer male white students, criminal conviction systems do the same and there are several examples like this.

B3) Facebook runs a multi-million dollars company based on target advertisements given to its users based on the data collected to them. The advertising system allows the ad publishers to specifically target their ads to certain groups of individuals that are identified by Facebook company. Subsequently the ads are delivered by the advertising platform.
The HUD , an entity for fair housing in USA, opened a case against facebook for unlawful housing practices. In particular , it is not legal in USA to give advantage to certain groups of people over others based on protected attributes such as gender,ethnicity, etc .. However, the factual allegations against facebook are that it allowed its ad publishers to target ads thus creating an inequality in the housing market and disadvantaging certain groups of people.

The HUD required facebook to pay a civil penaly, compensate all the affected people and train its employees on anti-discriminatory practices.

One of the main arguments of Facebook is that it only delivers the ads. However, there is no way in Facebook to broadcast an advertisement. In addition, some studies have shown that based on the characteristics of the advertisement Facebook targets the delivery of the ad (in the phase where the ad publisher has no longer control over the ad) therefore asuming the responsibility of doing so in a sensible market as the housing market.

d) Assistance obtained in previous years : This variable could harmful because it would deny help to people that need it that received help in the past years could belong to profes labour market insertion might be particularly difficult.

2)

A better goal would be to measure effectiveness of the program amount of people in difficult situation helped and identify those based on data, require more help for finding a job. For defining t situation it is enough to keep as measure the age group, the assi and occupational group. Definitely the gender feature should be is not ethical and not legal.
This goal would also change the perspective for treating people f category that need special support for being inserted into the job measure like the one described before, of excluding them of help the demographic disparity already present. In contrast, with our feedback loop is opened and the retroaction can be defined as " our system is helping to reduce the disparity.

B. THEORY (TOT 7,5 points). Answer briefly to the questions:

The Algorithmic Accountability Act:

i. Summarize the main points (no need to go into the details of numbers) (5 points)
ii. Briefly comment on the possible use of algorithmic fairness criteria (independence, separation, sufficiency) and/or bias measures (Gini and Shannon index) for the prescriptions of the law proposal (2,5 points).

The algorithmic accountability act was a bill passed in the US senate (not approved) that had as a goal to make companies developing Automated Decision Systems (ADS) accountable for it by performing an impact assessment of the system.

An ADS is defined as any computational process (included Machine Learning) that takes a decision or helps to take it. An example of it can be the COMPAS system used in the US for helping judges in taking a decision about the sentence of a defendant based on the probability of revidicism (given as a score from 1-10). In tha Act it is also defined the concept of a high-risk ADS , which is a system that in this context is used over particularly delicate decisions or by public entities.

The Accountability Act requires any company developing those systems to perform and imapct assessment to check that the ADS does not have any bias and does not systematically discriminate against a any group of people.

ii) The fairness criteria are a series of measures about the results of an algorithm that aim to measuring the fairness of a given algorithm, which means it gives predictions independently of a protected attribute like gender,ethnicity, ...
There are several existing measures. Firsto of all we have the independence :
P(R=r|A=a)=P(R=r|A=b)
Separation:
P(R=r|Y=r and A=a) = P(R=r | Y=r and A=b)
Sufficiency:
P(Y=r|R=r and A=a) = P(Y=r | R=r and A=b)

In addition balance meaures could be used to check that the datasets used to train the algorithms are not imbalanced. Those are measures that aim to have heterogeneous dataset. Therefore, those datasets correctly represent all classes.

The law proposal could include the requirement for companies to evaluate the ADS based on this fairness criteria in such a way that no systematic discriminatios is done against any group of people based on a protected attribute.

Describe the business of clearbox

Clearbox AI is a company that aims at applying the data ethics principles into the AI technologies. It is a company that offer synthetic dataset generation to help companies to preserve data privacy, improve data quality and automate testing in AI and analytic projects.

The usage of synthetic data is proposed as a solution for the issues regarding the privacy with respect to sensitive data. The analysis of sensitive data is crucial but the privacy principles must be respected. However, synthetic data pose a simple solution for the anonymization since it improves the de-identification of the dataset and creates a dataset that can be shared without criticality.

Secondly, the data quality of the dataset might be improved by mitigating the data scarcity and solving ML problems by solving balance problems. It can be further used to fix biases present in the dataset.

In order to create a good synthetic dataset and not allow identification disclosure the Gower's distance metric is used to measure the distances between datasets. It is 0 when a record in the synthetic dataset is also present in the original one. It is also used the DCR (Distance to Closest Row) and if this metric is too low it means that there could be a potential re-identification. This process makes very difficult the re-identification of the datasets.

DESCRIBE THE HUD CASE AGAINST FACEBOOK. THE FACTUAL AND LEGAL ALLEGATIONS

Facebook is a multi-million dollars company that base its business in using the data about its millions of users to charge advertisers for deploying target advertisements. The process works with an ad publisher that chooses the specific characteristics of the ad to publish and pays a fee to facebook for its service. After the advertiser has choosen the audience to which he desires the ad to be shown Facebook deploys the ad.

In particular there are housing advertisements published on Facebook. The HUD case is the case presented because of the unfair practices done by Facebook in this context.

Recall that Facebook defends itself from those claims by stating that it does not choose the audiences to which the ad is shown and this is done entirely by the ad publisher in the ad creation phase. However, scientifically studies have shown that Facebook do target the ad in the deployment phase based on the content and characteristics of the ad (budget, image, ...) using ,very likely, a ML algorithm so Facebook is also accountable by the ad targeting and not only the ad publisher. Those arguments are presented in the factual allegations of the case.

Secondly, we have to consider the USA context in which it is not allowed by the law to discriminate against any racial group (or any group of any protected attribute like gender, ethnicity, religion, ...) when publishing housing offers and those offers should be accesible to everyone. The legal allegation done by the HUD against Facebook is that in its platform housing ads were targeted and this process unfairly discriminated against some racial groups.

The HUD requires Facebook to pay a civil penalty for unfair housing practices and to compensate to every person affected by it. In addition, it is required to Facebook to train its employees on fair housing practices and to take action to not discriminate against racial groups in this context.

The case is still open in a court in the USA.

DESCRIBE THE ACM CODE OF CONDUCT

Describ

The Association of Computing Machinery has a code in which they describe the ethical obligations of software developers and leaders. This Code of Conduct aims to give guidelines and does not aim to be an algorithm for solving ethical issues.

General ethical principles

The code introduces the principles that each professional should guard.

a) The public good is the primary interest of any project. Any project should aim to be of good for society in general and whenever possible reduce the negative effects it might have. The code is very insistent with this point in the sense that if a system has the risk of being harmful even the possibility of not deploying it must be considered
b) Preserve the right of privacy of the people. This point of the code l sparticularly important in the context of Machine Learning where any information might be used for training an AI and obtain better results. However, it is necessary to do it respecting the rights of the individual also with respect to privacy.
c) Be fair and take action. Do not deploy systems that systematically discriminate against any group of people and take action in preventing it and correcting when necessary.

Professional Responsabilities

Each professional has a responsibility to maintain high ethical standards and to ensure that the developed systems comply with the general ethical principles. Be aware of the context in which the system will be deployed and be able to do a reflective analysis
Ensure that the systems are correctly evaluated for possible harmful effects.

Leadership Resposnabilities

Recognize the systems that are part of the society infrastructure and ensure that they act on the public good.

The code must be promoted and respected. The future of computing depends on technical and ethical excellence.

potentially be
. In fact, people
sions for which

based on the
professions that,
e difficult
stance received
removed since it

alling in A2
o market. In fact a
; only increases
solution the
negative" , thus

1) billed amounts : Clear proxy variable that discriminates against not wealthy people.

2) As said in the previous paragraph the model is based on a questionable assumptiuon that the people that need the most additional healthcare resources are the ones that have spent more money on it. Unfortunately, the world has the demographic disparity descreibed above so this assumption is trivially false. A possible change would be to eliminate this assumption and try to give the healthcare to the people that need it the most. In fact, people that would benefit the most are not wealthy. All variables correlated with financial status should be deleted, and where this is not possible (like with the appointments with specialist made) compensate the fact that this could lead to the increase of the disparity.

VARIABLES

As a context we must state as hyporthesis that economical status is linked to the ethnicity. In fact, in the US there is a historical demographic disparity in terms of economical status between white and other ethnicities. Hisctrically black and latinos have been disadvataged and population of immigrants that, unfortunately, do not enjoy of the same economical status as white people.

Admit code : The type of admission code. This factor does not seem to be linked with ethnicity.
First Generation Indicator : This indicator is higly linked with the economical status of the family. In fact, richest families will have parents with at least a Bachelor Degree. This factor is linked then with the ethnicity.
Median Income by Admission ZIP code: This factor is clearly linked with the economical status of the family and as a consequence to the ethnicity.
High School GPA: Altough this factor depends on an individual level, students that have not a peaceful environment at home have more probability of low academic performance. This attribute should be treated with specific caution because it is a good predictor.
Legacy Indicator: As the first generation indicator this factor is linked with the economical status of a family thus to their ethnicity.
International Indicator: Factor linked to ethnicity of student
High School Percentile: As the school GPA this could disavatage students in bad economical situation that do not have time for studying and belong to those categories.
High School Size: The size of the school is linked with the economical capacity of the family. As a tendency small schools are supposed to have better quality and have higher taxes.
Veteran Indicator: Is not linked with ethnicity (with available data)
In State Resident Indicator : It is not possible to know if it is linked with the ethnicity

2)
As discussed in the previous point the system is measuring the economical status of the student and giving an outcome based on it. It can be argued that this process is not ethic since it is only reinforcing the already present demographic disparity by closing the loop of demographic disparity.
Measures like legacy indicator and first generation indicator are highly segreative and there is the risk of discrimination against this type of students using an unfair argument. In fact, it is not under the control of student if his parents went to the university or not.

3)
The system using the already present variables is very good at predicting which students need additional support for continuing their studies and students in those situation would benefit from being identified and helped by the university. This action would create a "negative retroaction" in the feedback loop of demographic disparity and help to decrease the historical gap between the two populations.

However, if it is not possible to remove the bias of the system it is enough to delete all the variables that are linked with the economical status of the student and as a consequence to his ethnicity. The most problematic variables are the legacy indicator and the first generator indicator since take a decision based in a totally unfair argument.
Variables like GPA and percentile should be kept being cautios about the potential bias that can be present agaianst students and this bias should be compensated.

2)
The algorithm does not seem to preserve good the right of the citizens to have a privacy. While it could be argued that it is a proportional measure taken by the state to avoid tax evasion so it respects the balance principle it is not clear if the users gave the consent for their data to be processed in this way by the websites. Therefore, this measure could lead to a legal issue (GDPR) and also arise ethical issues about how far can the state try to "observe" its citizens. Additionally, as said before , there is the flawless assumption about the same online behavior of different ethnic groups which does not consider diversity of behaviors and penalizes people that uses this portals.
By Campbell's law the observation that the usage of those websites leads to further investigation of the personal finances by the state could lead people to simply stop using those websites and use other online e-commerce sites that are not affiliated with French government. Finally, this can lead to e-commerce websites to not cooperating with the state because of the loss of customers.

the text, w
there must

Before ana
some histo
groups (bla
addition, w
them even
Italian citiz
No informa
could be ta

VARIABLES
a) Birth
the c
b) Age
How
In fa
smal
c) City
som
d) Car
ethn
e) Clain
f) Year
Italia

2)
The measu
population
a) The
for t
b) Age
loos
It co
c) Clain
time
d) Year
pens
3)Recall th
P(R=1|A=a
Therefore
example, i

Describe th

The Associ
in computi
However, t
guidance.

PRINCIPLES
The Code o

1) Any
bene
2) No p
3) Peop
qual
impa

RESPOSNA
1) Ensu

LEADERSH
It is necess

as not used for training the classifier. Therefore, we conclude among the used predictors
be at least one variable with a correlation with the ethnic group.

lyzing in detail the variables we can understand the context of the Italian society. There are
rical demographic disparities affecting mostly immigrant population. Usually those ethnic
ck and asiatic) have less economical resources than nationals (usually white people).In
e could hypothize that people that have been living less time in Italy have less claims against
f they had an accident in other country. This situation could lead to the penalization to the
ens (mostly caucasian).
tion is available about the training set. A balance test must be done in it (and this hypothesis
sted)

ANALYSIS
Place : This vairbale is strongly connected with the ethnic group and could be the source of
iscrimination done by the algorithm
With the information on the text it is not possible to determine if there is any correlation.
ever, we can observe that italian population (white) is generally older than in other countries.
t, Italy is one of the countries where population is aging the fastest. So, we could hypothize a
correlation between age and ethnic group.
The city of residence might be correlated. Immigran population could concentrate more in
urban areas than others. However, there is is not enough information.
The type of the vehicle is correlated to the economical status of the person, thus to his/her
city.
history : This variable might be lower for young and immigrants.
distance : This variable might be correlated with economical status in the sense that among
n population the richest take the less the car (live closer to the center of the city).

res taken have a series of issues that do not take into consideration the variety of the
, in particular:
birthplace of a person does not seem to be correlated with the incident rate. It might be only
e possibility of newly arrived foreigns. However, this could be measured in other way.
The age could be related with the probability of an accident given that elder people tend to
the capacity of reaction. However, it must be used careful for not over-penalize old people.
ld be used to compensate for the claim-history as we will see below.
history : This variable could incorrectly penalize elder people that have had the license more
it could also benefit the foreign people. it must be normalized by the age of the person.
distance : This variable might be correlated with the probability of an incident but unfairly
also people working "on the street" rather than on an office.
t separation is a fairness criteria based on the fact that
and Y=1) == P(R=1|A=b and Y=1).
for being able to compute it we must ask participants wether they have had an accident , for
the last year.

e ACM Code

ation of Computing machinery has a code an ethical code of conduct to give people working
g machinery (software development, machine learning, ...) a guide for ethical issues.
his code must not be used as an algorithm to solve ethical problems but as a tool for

s
describe a series of principles that should be respected.

oftware developing must be done for the public good. As public good it is intended that it is
ficial for humankind.
erson should be discriminated by the software
e have the right to privacy. This is particularly important in the context of ML in which huge
tities of data must be used to train algorithms and often data scientists forget about the
rtance of respecting people privacy.

BILITES
re and take action against any systematic discrimination done against any person

P RESPONSABILITIES
ary to ensure that the product is good and respects all ethical