

ETHICS IN ARTIFICIAL INTELLIGENCE

Module 1

1) Introduction to ethics

1.1) Morality

There is no a univocal definition of morality, but in a general sense it refers to a set of rules that we can use to distinguish between right and wrong actions.

1.1.1) CONVENTIONAL AND CRITICAL MORALITY

Conventional morality = rules and principles shared by a particular society, they can vary over the space and time. Usually we are agree with the moral values of the society where we was born into. → CULTURE DEPENDENT

Critical morality = it represents standard moral values, shared by all the people independently by the society they come from (it doesn't depend by cultural factors). Sort of metric to evaluate the conventional morality → CULTURE INDEPENDENT

1.1.2) BRANCHES OF MORAL PHILOSOPHY

Value theory = field that studies values (human values, it answers to question like what is a good life, what is happiness etc)

Normative ethics = branch of moral philosophy, it establishes rules to define how people have to act in some situations in order to be moral. (What does it make an action right? Which is the best behavior in that situation etc...). So it defines what is right and how we have to behave based on moral principles. → JUDGMENTS ABOUT OUR BEHAVIOUR

Non normative ethics = it's not focused on the people behaviour, but it's focused in how we can describe the moral behaviour, it gives us definitions in order to understand if our actions are moral. We can distinguish 2 areas:

1. Descriptive ethics: it studies how people think about morality, it uses empirical method to study how and what people choose in different cases but it doesn't evaluate our actions (it studies different concept of morality for different countries for example)
2. Meta ethics: field that analysis the language, concepts and methods of reasoning in normative ethics. → ANALYSIS OF OUR BEHAVIOUR

1.1.3) MORALITY VS OTHER NORMATIVE SYSTEMS

Ethics is not always aligned with other normative systems.

Laws = laws can be non moral, some legal actions can be unethical (like betraying (tradire) or cheating) and some non legal actions can be ethical (like criticising a dictator).

Etiquette (buone maniere) = they are a different concept, we can't define these principles ethical or unethical (galateo is an etiquette but I can't define myself unethical if I don't follow that set of rules).

Personal interests = sometimes if we want to be moral we have to sacrifice our personal interests, and sometimes could happen that we can obtain advantages in an unethical behaviour, this represents a conflict between the concept of ethics and personal interests.

Traditions: they couldn't be moral even if they are respected for a long time.

Religion: the relation between morality and religion is not simple, it's not clear if God commands something because it's morally right or actions are moral because God impose them to us.

1.1.4) ABSOLUTISM AND RELATIVISM

Absolutism = single true ethics.

Relativism = judgments are relative to particular frameworks and attitudes.

1.2) Consequentialism → FOCUS ON THE CONSEQUENCES

Optimific action = action that produces the best overall results (best consequence).

Consequentialism = family of theories that consider an action morally if and only if it's optimistic.

→ CONSEQUENTIALISM → Action is moral if it generates benefits.

1.2.1) ACT UTILITARIANISM

Act utilitarianism = instance of consequentialism that uses as metric to evaluate the morality the well being generated by an action (here we evaluate each action)

Principle of utility = moral standard of act utilitarianism that consider an action moral if and only if it maximises the well being generated with respect to all the other possibilities

Pros:

1. All the people are equally important (it's important the well being of everyone).
2. It's coherent to some basic moral intuitions (for example slavery is not optimific).
3. It eliminates ambiguity, the unique goal is to maximize the wellbeing.
4. Flexibility, even if an action is unmoral e can do it anyway if it increases the well being of everyone.

Benefit generated by a single action.
It should be the best action over all the possibilities

Cons:

1. Sometimes we have to sacrifice ourself for the common wellbeing. Sometimes we could be impartial (we only consider our well being)
2. There is no standard in the way to take a decision.
3. If I'm part of a minority, my well being could be less relevant because it interests a small portion with respect to all the people

1.2.2) RULE UTILITARIANISM

Rule utilitarianism = a rule is morally right if it's required by a optimific social rule.

→ RULE UTILITARIANISM = Benefits can be produced if all the people act in that way

Optimific social rule = rules that, if all the people decide to follow them, could achieve the best results for the society in terms of well being (here we evaluate a set of rules).

Note: difference between ACT and RULE

Action Utilitarianism judges each action by its outcome: if lying in one case leads to more happiness, then it is right to lie.

Rule Utilitarianism, on the other hand, says: even if lying in that case leads to more happiness, if everyone lied often, trust in society would collapse, so the rule "don't lie" is preferable — and that's the one to follow.

1.3) Deontology → FOCUS ONLY ON THE MORAL RULES

Deontology = an action is morally right or wrong based on some moral rules, we don't care about the consequences (it doesn't matter the result of the action, but only if it respect the moral rules).

Example: Telling the truth is morally right, even if in some cases telling the truth causes pain to someone. Lying is morally wrong, even if in some cases lying would avoid problems. Deontology says we have to do what is right.

1.3.1) KANTIAN ETHICS

Kantian ethics = ethical theory based on the fact that rules hold for everyone in the same way (fairness and consistency = lealtà e coerenza)

Maxim = personal rule that describes why we perform a certain action in a certain situation (we are interested to the intentions instead of the results). → NO INTERESTED IN CONSEQUENCES
→ UNIVERSALIZABLE if my goal is reachable in case all the people act like me

Universalizable maxim = a maxim is universalizable if all the people perform the maxim action in the maxim situation (the maxim could be shared by all the people) and the goal of my action is achieved in such a word. If we can reach the goal when all the people perform that action, it means that our intentions are not egoistic, viceversa if our goal is not reachable it means that our intentions are immoral. For example, promising something even if we already know that we are lying: if all the people follow this behaviour, the trust in the promises would disappear.

Principle of universalizability = an action is acceptable if and only if its maxim is universalizable. If we act immorally we are irrational (we know that we can't do that, but we do it anyway)

Hypothetical imperatives = imperatives that require us to do what is needed to reach our goal (if we want to reach this result, we have to do that). They depends by our needs, it is not correlated to the concept of morality (if you want to pass the exam, you have to study).

Categorical imperatives = Imperatives that don't depend on a single individual but are applicable to all the people, disregarding them make us irrational (there is no goal behind this rules, we should follow them because they are morally right, for example don't be a layer or don't be a killer).

Critics = some principles could be shared (so universalizable maxim) but at the same time they could be immoral (for example racism). The problem is that for Kant what matters is living following shared principles, but we can't be sure that these principles are morally right.

1.3.2) DAVID ROSS'S PRIMA FACIE DUTIES

Prima facie duties are non-absolute moral obligations.

- They are valid unless they conflict with a stronger duty (for example I can promise to a friend of mine that we will study together, but if while I go to his home I find someone that need medical help I can don't respect the promise made before).
- According to Ross, morality is made of practical judgment, common sense and moral sensitivity.
- It is a flexible deontological theory, which combines moral rules and concrete context.

1.4) Proceduralism

→ FOCUS ON A PROCEDURE

Proceduralism is an approach to ethics that does not start from fixed moral principles, but instead says:

"We follow a just procedure to determine which moral principles are valid."

In other words:

- We do not assume from the outset what is right or wrong.
- We establish a rational, impartial, and just procedure, and we accept as moral what results from this procedure.

1.4.1) CONTRACTARIANISM

Contractarianism (political) = political theory which states that laws are right if and only if they would be accepted by free, equal and rational people (each one limiting self-interest = rational)

Contractarianism (moral) = actions are morally right if and only if they would be accepted by free, equal and rational people

State of nature = situation where there is no government, it means that each individual acts to maximize its own self-interest.

However, the effect is that everyone will be in worse conditions, to improve the situation cooperation among people and a government are required.

Contractarianism characteristics: in general the morality is a social phenomenon since moral rules are decided through

cooperation (we want to escape by the state of nature). The procedure to decide the laws is seen before, it consists in asking if free, equal and rational people would agree to a certain rule. Moral rules can be violated if people stop to cooperate.

→ To decide if an action is moral, each person into a group of people should be agree with a certain common decision

1.5) Virtue ethics

Family of theories that considers an action morally right if and only if a virtuous person (i.e., an ideal character, a role model) would do.

Problems = it could be too demanding (esigente) since it based on the virtues, there are no absolute rules to follow (different virtuous men could have different virtues), there are no rules to deal with virtues conflicts.

1.6) Principlism

→ FOCUSED ON 4 PRINCIPLES

Principlism = it's an ethical approach that relies on the application of four fundamental principles to resolve moral dilemmas. Instead of following a single ethical theory (like utilitarianism or Kant), principlism uses multiple basic moral principles, which are balanced on a case-by-case basis. The four fundamental principles (according to Beauchamp and Childress):

1. **Beneficence:** Promoting the good and well-being of others.
2. **Non-maleficence:** Avoiding causing harm ("primum non nocere").
3. **Autonomy:** Respecting the freedom of choice and the right to self-determination of the person.
4. **Justice:** Treating people fairly, distributing resources and benefits fairly.

Independently by the situation and subjects, these principles should be guaranteed

Moral Structure of Principlism = principlism distinguishes two major levels of morality:

1. **Common Morality** -> Moral norms shared by all rational people, they derive from universal human experience. They are absolute (e.g. do not kill innocents, help those who suffer, basic ethical foundation).
2. **Particular Morality** -> more specific or contextual moral norms, professional Morality (norms specific to a profession e.g. medical ethics), Public Policy like laws, regulations, guidelines of public institutions (They must respect common morality), they can vary based on culture, context, role.

2) Trustworthy AI in the EU → AI affidabile

2.1) AI4People's Ethical Framework for a Good AI Society

Forum created in 2018 with the goal to define principles for a good AI society.

2.1.1) OPPORTUNITIES AND RISKS OF AI FOR SOCIETY

Opportunities = it can help humans improving cultural, intellectual and social activities, can support human decision making, it can help in solving problems.

Risks = humans shouldn't devalue human skills, it shouldn't lift humans from responsibilities, it must be supervised by humans.

2.1.2) UNIFIED FRAMEWORK OF PRINCIPLES FOR AI IN SOCIETY

Common principles used by different organisations and initiatives:

1. **Beneficence** (ai must help humans)
2. **Non maleficence** (ai system shouldn't cause harm = danni)
3. **Autonomy** (balance between decision making delegated to AI and human)
4. **Justice** (AI must contribute to equality and global justice)
5. **Explicability** (ai should be understandable in the way it works)

2.1.3) RECOMMENDATIONS FOR A GOOD AI SOCIETY

1. **Assessment** (valutazioni) -> evaluate that the current regulations are able to guaranteed an ethical usage of the AI, evaluate what we can delegate to an AI, evaluate possible risks.
2. **Development** -> develop a framework to enhance (migliorare) explicability, procedure to evaluate AI decisions, the quality of AI and the supervision of possible AI mistakes and problem generated by it.
3. **Incentivization** -> incentivize research, ai socially preferable, ethical legal and social considerations of this phenomenon.
4. **Support** -> support companies to understand the ethical implications of their products and the creations of dice of conducts for data and ai professions.

2.2) AI HLEG's Ethics Guidelines for Trustworthy AI

Group established by the European Commission in 2018. Its goal is defining guidelines for AI ethics and defining policy and investment recommendations.

Ethics guidelines for trustworthy AI = published in 2019, it represents an ethic framework based on these 3 principles:

1. **Lawful** -> ai must adhere to laws and regulations like EU treaties and fundamental rights, gdpr, international treaties etc
2. **Ethical** -> ai must be aligned with ethical principles and values (important where there is no laws yet)
3. **Robust** -> ai must be socially and technically robust in order to minimize harms (danni).

The document is composed by 3 chapters:

1. **Foundations of trustworthy AI** -> principles that an AI should respect
2. **Realisation of trustworthy AI** -> requirements to achieve trustworthiness
3. **Assessment of trustworthy AI** -> trustworthiness assessment methods

2.2.1) CHAPTER 1: FOUNDATIONS OF TRUSTWORTHY AI

It's based on fundamental rights described in the EU treaties and international human rights laws (for example respect for human dignity, human freedom, equality among citizens and respect for their rights).

Principle of respect for human autonomy = AI can't manipulate the behaviour of humans, we always have to be able to decide what really want to do.

Principle of prevention of harm = AI must be safety and robust, attention must be paid to vulnerable people.

Principle of fairness (equità, lealtà) = it must be present in a substantive dimension (the results must be unbiased, everyone can use technology etc) and in a procedural dimension (we have to be able to contest and correct what the AI system infers if we think that it made a mistake).

Principle of explicability = AI system decisions must be as most explainable and clear as possible.

2.2.2) CHAPTER 2: REALIZATION OF TRUSTWORTHY AI

Concrete requirements form the principles of the previous chapter.

Who is involved = developers (who research and develop AI system), deployers (who use AI system in its business) and end users (who use the final AI system)

The main requirements are:

Human agency and oversight (supervisione) = AI system just should enhance (rafforzare) autonomy and decision making. We have to evaluate all possible risks, users must be informed in order to understand how to use AI systems in a responsible way and they have the right to not be subject to only automatic decisions if this affects them. There are several methods to **oversight mechanisms**: human in the loop (humans intervention in every decision), human on the loop (human intervention in the design cycle and monitoring of the system's operation) and human in command (human decides if, how and when using AI systems).

↳ AI should be helpful to people. Need for oversight methods

Technical robustness and safety = measure to prevent and minimize unintentional harms. AI system must be protected against attacks that target (mirare a) the data and the model, there should be a fallback plans in case of problems and an explicit evaluation process, the output of an AI system should be reliable and reproducible.

Privacy and data governance = quality and security of data should be guaranteed, so data provided by the users should be protected, datasets should be cleared from biases and inaccuracies.

Transparency = all the elements of the AI system must be transparent, a documentation of the system is needed, AI decisions must be understandable, AI limitations should be communicated.

Diversity, non discrimination and fairness = inclusion and diversity should be considered, biases should be removed (oversight datasets), AI systems should be designed in order to be used by all the people without discrimination.

Societal and environmental well being = it's important to keep into account how an AI system could impact on the users (in terms of people's physical and mental well being)

Accountability (responsabilità) = clear responsibilities should be defined for decisions made by AI systems, so we have to evaluate the AI systems in order to identify possible negative impacts and prevent them, moreover a redress (risarcimento) mechanism should be present.

Technical methods to ensure trustworthy = embed trustworthiness requirements into the AI system as constraints, technique to explain mechanisms, defining tests and validate system during its lifecycle.

Non technical methods = introduction regulations, educate and train enrolled stakeholders, be transparent etc

2.2.3) CHAPTER 3: ASSESSMENT OF TRUSTWORTHY AI

The chapter defines a generic assessment (valutazione) list to implement the requirements of chapter 2. It was first tested by a small group of companies and then extended to all the other stakeholders in order to collect feedbacks.

Assessment list = steps to assess the trustworthiness of an AI system.

1. it should be specific based on the use case
2. It can be integrated into existing governance mechanism
3. It's continuously improved

3) Human rights

Human rights = they are ethical principles that should be guaranteed for all the people.

Negative liberties = rights where third parts don't interfere (for example, the right to speak means that no one can impose me to can't tell what I want)

Positive liberties = rights that I ask to be guaranteed by a third part (for example, right to education must be guaranteed by the state)

Module 2

1) AI in the GDPR

GDPR is the General Data Protection Regulation. It is a European Union law that came into force on May 25, 2018, with the aim of protecting the personal data of European citizens.

The GDPR establishes how companies, public bodies and organizations must collect, process and store people's personal data.

1.1) Introduction

1.1.1) DEFINITIONS (ARTICLE 4)

Personal data = any information related to an identified or identifiable (if we can identify it directly or indirectly using some features) natural person (not companies, which are legal person, but only individuals).

Processing = any operation performed on personal data

Controller = who decides the purposes and how to process personal data.

Processor = who processes personal data on behalf (per conto) of a controller.

1.1.2) TERRITORIAL SCOPE (ARTICLE 3)

The GDPR applies to the processing of personal data whenever:

1. The controller or processor resides in EU
2. The data subject is in the EU when the purpose is for offering goods or services and monitoring of behaviour.

1.2) Data protection principles

1.2.1) LAWFULNESS OF PROCESSING (ARTICLE 6)

Processing of personal data must respect at least one of the following conditions to be defined lawful:

1. **Consent** = the data subject has given consent to process its personal data
2. **Necessity** = personal data is necessary for certain aim (protect vital interests, public interest, contract for an insurance etc)
3. **Legitimate interest** = legitimate interests of the controller if and only if they don't override (non prevalgono) the fundamental rights of the data subject

1.2.2) TRANSPARENCY (ARTICLE 5)

Any information regarding data processing must be clear and accessible

1.2.3) FAIRNESS (ARTICLE 5)

Data subjects should be informed of the existence of data processing (informational fairness), controllers should implement measures to correct inaccuracies and protect data (substantive fairness)

1.2.4) PURPOSE LIMITATION (ARTICLE 5)

Personal data can be collected for specified and legitimate purposes. Purpose always legit are public interest, historical or scientific research and statistical.

1.2.5) DATA MINIMIZATION (ARTICLE 5)

Data collected from the data subject should be adequate and limited with respect to the purpose it is required for.

1.2.6) ACCURACY (ARTICLE 5)

Personal data should be accurate and eventually updated if necessary. Inaccuracies must be rectified.

1.2.7) STORAGE LIMITATION (ARTICLE 5)

Personal data should be kept only for the time needed for its purpose.

1.3) Personal data (article 4.1)

1.3.1) IDENTIFIABILITY

Identifiability = condition under which some data not explicitly linked to a person allows to still identify that person (they are still considered personal data)

Pseudonymization = substitute data items identifying a person with pseudonyms (link between person and pseudonym can be tracked back)

Anonymization = substitute data items identifying a person with non linkable information

1.3.2) INFERRED DATA

Inferred personal data = new information about a data subject obtained using algorithmic models on its personal data.

Right to “reasonable inference” = decisions made by AI systems must be based on reasonable inferences, that is, on logical, fair, understandable and justified conclusions from the personal data collected (conclusions can be unreasonable if they don't affect data subjects). The criteria to define a reasonable conclusion are the following: **acceptability** (input data should be coherent and acceptable for the final purpose, ethnicity cannot be used infer whether an individual is a criminal), **relevance** (the inferred information should be acceptable for their final purpose, ethnicity cannot be inferred from the available data if the purpose is for approving a loan) and **reliability** (data and processing methods should be curate and statistically reliable).

1.4) Profiling (article 4.2)

Profiling = system that predicts the probability that an individual having a feature F2 knowing before that it has a feature F1.

So, if you has F1, which is the probability that you has also F2?

In GDPR is defined as any form of processing of personal data that produces legal effects or significantly affects it.

According to the European data protection board profiling is the process of classifying individuals (or groups) based on their features.

Note: Cambridge Analytica scandal

The Cambridge Analytica scandal broke in 2018, when it was revealed that a company called Cambridge Analytica had obtained the personal data of around 87 million Facebook users (without their consent) and used it to influence political opinions.

1. How did they get the data? A psychology researcher (Aleksandr Kogan) created a Facebook app called: "thisisyourdigitallife". The app looked like a simple personality quiz, but it requested access to: The user's personal data And also the data of their Facebook friends (without those friends knowing!)
2. About 270,000 people used the app, but through them, the data of tens of millions of users was collected.
3. This data was then sold to Cambridge Analytica, a company working in political consulting and data analysis.

What were the data used for? Cambridge Analytica used the data to:

- Create psychological profiles of voters
- Target people with personalized political ads, such as:
- Tailored content to influence someone to vote for a specific candidate
- Or to discourage certain groups from voting at all

In which campaigns was the data used?

- Brexit (UK, 2016): to support the campaign for leaving the EU
- US presidential election (2016): to support Donald Trump's campaign

1.4.1) SURVEILLANCE

Surveillance capitalism = capitalistic system where human experience and behaviour become marketable entities, they could be defined by an economic value.

Surveillance state = system where the government uses surveillance, data collection and analysis to identify problems, govern population and deliver social services.

1.4.2) DIFFERENTIAL INFERENCE

Differential inference = make different predictions depending on the input features. It could be very useful but at the same time dangerous in case discriminative features are used to solve certain tasks (for example, using ML to predict health issues provides benefits to all the data subject, but ML with health data for recruiting would worsen the situation of who is already disadvantaged).

1.4.3) DISCRIMINATION

There are 2 main opinions on AI systems:

1. AI can avoid fallacies of human psychology
2. AI can make mistake and discriminate

Direct discrimination = when AI system bases its predictions on protected features

Indirect discrimination = when AI system has a disproportional impact on a protected group without a (understandable) reason. Since AI could be trained on datasets that reproduce past human behaviour or judgment, or could be trained on biased datasets, even if the training is supervised we could obtain a discriminative behaviour anyway.

1.5) Consent (article 4.11)

Consent = explicit agreement of the data subject that allows to process its personal data. It should be **freely given** (data subject has to choose if giving consent for profiling), **specific** (data should be used for a specific and compatible purpose), **informed** (data subject should be clearly informed) and **unambiguously provided** (consent should be explicitly provided).

1.5.1) CONDITIONS FOR CONSET (ARTICLE 7)

Some requirements for consent are:

1. Controller must be able to demonstrate that the data subject has provided its consent.
2. The consent should be clearly provided.
3. Data subject has the right to withdraw its consent at any time.
4. Consent must be assessed if the performance of a contract is influenced on consenting the processing of personal data.
5. Consent is always no freely given if the subject data risks in case it refuses to consent.

1.6) Data subjects' rights

1.6.1) CONTROLLERS' INFORMATION DUTIES (obblighi informativi) (ARTICLE 13-14)

When personal data is collected, the controller should provide to the data subject the following informations:

1. The identity of the controller
2. The contact details of the data officer (responsabile dei dati)
3. Informations on the data: purpose and legal basis of the processing, categories of data collected, who will use these data, period of time how long the data is stored
4. Possibility to retire the data and to lodge a complaint (fare un reclamo)
5. communicate if the data will be used for profiling (in that case controllers has to specify more details about the profiling task, ike the possible consequences, the overall purpose of the system and how it works).

1.6.2) RIGHT TO ACCESS (ARTICLE 15)

Data subjects have the right to be informed about the processing of their data and to access both input and inferred personal data

1.6.3) RIGHT TO RECTIFICATION

Data subjects have the right to rectify their personal data. In general, data can be rectified when the correctness can be objectively determined and when the inferred data is probabilistic and there was a mistake during inference.

1.6.4) RIGHT TO ERASURE (cancellazione) (ARTICLE 17)

Data subjects have the right to have their own personal data erased without delay from the controller when:

1. Data is no longer necessary for the purpose it was collected for
2. Data subject withdraws its consent (unless there are other legal basis)
3. Data unlawfully processed
4. Data have to be erased for legal obligations

In some cases this right is not apply, for example for public interest in healthcare, scientific or statistical purpose, for legal and defence claims etc.

1.6.5) RIGHT TO PORTABILITY (ARTICLE 20)

The right to data portability allows you to:

1. Obtain your personal data from an organization (data controller),
2. in a machine-readable format (e.g. CSV, JSON),
3. and easily transfer it to another organization.

1.6.6) RIGHT TO OBJECT (diritto di opporsi) (ARTICLE 21)

Data subjects have the right to request the termination of the processing of their data when all the following conditions are met:

1. Data subject has reasons to withdraw
2. The reason for processing is public or legitimate interest
3. Controller cannot demonstrate legitimate interest for processing data.

1.6.7) RIGHTS WITH AUTOMATED DECISION-MAKING (ARTICLE 22)

Data subjects has the right to not have decisions based only on automated profiling if it produces illegal effects or significant ones.

Exceptions are applied when data is needed to enter or perform a contract (allowed to use automated system to process a high number of job applications), when authorization is given by the authorities and when explicit consent is given.

1.6.8) EXPLAINABILITY IN THE GDPR (ARTICLE 22, RECITAL 71)

It's not clear if the GDPR considers the right to explanation an obligation of the controller, due to the fact that recital 71 mentions the right to an explanation while article 22 doesn't.

1.7) Risk-based data protection

Risk based legislation = measures with the goals actively preventing risks.

*Design a safety
→ AI system*

1.7.1) DATA PROTECTION BY DESIGN AND BY DEFAULT (ARTICLE 25)

Controller must implement technical and organizational measures to respect data protection principles (protection by design), and it must ensure that only the necessary data is processed for each purpose (protection by default).

Working in an ethical way

1.7.2) IMPACT ASSESSMENT (ARTICLE 35-36)

Controllers must preventively perform impact assessment to processing systems that are likely to have high risks in terms of rights and freedoms of the data subjects. In case the risk is high, controller must consult the supervisory authority which will provide its written advice (parere scritto).

1.7.3) DATA PROTECTION OFFICERS (ARTICLE 37)

Controllers must appoint (nominare) a data protection officer to ensure compliance (conformità) with the GDPR if processing requires continuous monitoring on data subjects, involves large scale sensitive data or concern criminal convictions (condanne).

2) CLAUDETTE

Claudette = it's a clause detector, a system to classify clauses in terms of services or privacy policies as:

1. Clearly fair
2. Potentially unfair
3. Clearly unfair

Unfair contractual term = a contractual term that is not individually negotiated is considered unfair if it causes a significant unbalance in the parties' rights and obligations.

2.1) Unfairness categories

The following are unfairness situations:

Consent by using clause = potentially unfair if consumer accepts the terms of service by simply using the service.

Privacy included = potentially unfair if the consumer consents to the privacy policy by simply using the service.

Unilateral change = potentially unfair if the provider can unilaterally modify the terms of the service.

Jurisdiction clause = clearly unfair if it only allows judicial proceedings in a different city or country (viceversa is fair).

Choice of law = clearly fair if the law of the consumer's country of residence is applied in case of disputes (viceversa is potentially unfair).

Arbitration (arbitrato, procedura penale) clause = same of jurisdiction clause

Limitation of liability (responsabilità) = fair if the provider may be liable, potentially unfair if the provider is never liable unless obliged by law, clearly unfair if the provider is never liable

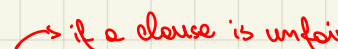
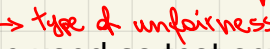
Unilater termination = if the provider has the right to suspend or terminate the service and the reason are specified (potentially unfair) or without any reasons (clearly unfair)

Content removal = same as before, but in this case related to user's content.

2.2) Methodology

Training data = manually annotated terms of service

Tasks = 2 kind of tasks are solved:

1. **Detection** (binary classification, sentence contains or not an unfair clause) 
2. **Sentence classification** (determining the category of unfair clause) 

Experimental setup = leave-one-out (we have n document, one will be used as test set and the others are used as training set (4/5) and validation set (1/5).

Metrics = precision, recall, f1

2.2.1) BASE CLAUSE CLASSIFIER

Methods like bag of words, tree kernels, cnn, svm...

2.2.2) BACKGROUND KNOWLEDGE INJECTION

Memory-augmented neural network = model that given a query, it retrieves some knowledge from the memory and combines them to produce the prediction. The Claudette's knowledge base is composed of all the possible rationales for which a clause can be unfair.

Workflow: 1) the clause is used to query (interrogate) the knowledge base using a similarity score and the most relevant rationale is extracted, 2) rationale and query are combined, 3) repeat until the similarity score is too low 4) at the end the model computes the prediction and provides the rationales as explanation.

2.2.3) MULTILINGUISM

Training data = same terms of service of the original Claudette corpus selected according to the following criteria: 1) the term of service is available in the target language, 2) there is a correspondence in terms of versions and structures similarities between the documents in the 2 languages.

Approaches = different approaches have been experimented (we go from the most precise and expensive to the lowest precise but cheapest)

1. **Retrain Claudette from scratch with newly data in the target language (with human annotation).** → **BEST ONE**
2. Use machine translation to translate annotate English document in the target language, match the machine translated document with the original one in the target language (human annotation), and train Claudette from scratch.
3. Training set translation (automated)
4. Machine translation of queries (the document is translated and we use the English version of Claudette, then the results are translated again)

2.3) CLAUDETTE and GDPR

Claudette for GDPR compliance (conformità) = to integrate Claudette as a tool to check GDPR compliance, three dimensions (each containing different categories) are checked:

1. **Comprehensiveness of information** = policy should contain all the information required by articles 13 and 14 of the GDPR.
2. **Substantive compliance** = policy should contain process personal data complying with the GDPR
3. **Clarity of expression** = policy should be precise and understandable (evaluation of the language used, no ambiguous or unclear sentences).

2.4) LLMs and privacy policies

Comprehensive policy from LLMs = let LLMs extract relevant information from privacy policies for comprehensiveness (policy should contain all the relevant information like categories of personal data collected, purpose each category is processed for, labels basis for processing each category etc).

Experimental setup = in general we can check questions like what data does the company process about me? For what purpose does the company use my email address? etc...

Three scenarios were considered:

1. Human evaluation of the questions on existing privacy policies.
2. LLMs to answer the questions on ideal mock policies (with human evaluation).
3. LLMs to answer the questions on real policies (with human evaluation).

Results show that:

1. LLMs have high performance in the mock policies.
2. LLMs and humans struggle to answer the questions on real privacy policies (they could be ambiguous or unclear)

3) DISCRIMINATION

Disparate treatment = the outcome of an algorithm is based on protected features (intentional discrimination, we use racial or gender feature to decide an assumption for example)

→ DIRECT DISCRIMINATION

Disparate impact = the outcome of an algorithm that uses neutral features is disproportionate against certain groups without an acceptable reason (indirect discrimination, neutral features correlate to protected ones).

→ INDIRECT DISCRIMINATION

3.1) Biased data

3.1.1) HISTORICAL BIAS

Historical bias = system trained on intrinsically biased data will reproduce the same biased behaviour. (Data can be biased if it comes from past human judgment)

3.1.2) PROXY VARIABLES

Proxy variables = neutral features that is connected to a protected one resulting in a disparate impact on a certain group (example: an automated system that decides which communities to send aid to after a disaster such as a flood or earthquake. The system makes its decisions based on data from past insurance claims. This seems objective, but in reality it is a proxy for a social factor: insurance claims often reflect people's economic conditions. In fact, poorer communities often cannot afford insurance or make fewer claims. As a result, the system sees less data about those areas. And therefore, it incorrectly concludes that they do not need help, even though they may be the hardest hit by the disaster. As a result, the system appears neutral, but in practice it disadvantages low-income communities).

3.1.3) BIASES EMBEDDED IN PREDICTORS

Bias embedded in predictors = system that uses favourable features that only a certain group has.

3.1.4) UNBALANCED SAMPLES

Unbalanced samples = the dataset doesn't reflect the statistical composition of the population. It means that the algorithm will be inaccurate towards minorities.

3.2) Algorithm choice

3.2.1) AGGREGATION BIAS PROBLEM

Aggregation bias problem = system that has good results overall, but with poor performance for specific groups.

3.2.2) DIFFERENT BASE RATES

Base rate / Prior probability = proportion of samples belonging to a certain class.

Fairness criteria = the main fairness criteria are the following:

1. **Statistical parity** = each group should have an equal proportion of positive and negative predictions
2. **Equality of opportunity** = the members sharing the same features between different groups should be treated equally
3. **Calibration** = the proportion of correct predictions should be equal for each class within each group
4. **Conditional use error** = the proportion of incorrect predictions should be equal for each class within each group
5. **Treatment equality** = the error ratio of positive and negative predictions should be equal across all groups

Handling different base rates = we can do the following:

1. **Do nothing** (accept that different groups are actually associated to different probabilities).
2. **Modify the threshold for everyone** (raise or decrease the threshold to diminish the favourable classification for everyone)
3. **Change the decision for everyone** (use different threshold depending on the group or alternative measures).

Note: with different base rates, it is impossible to achieve all fairness criteria through thresholding.

4) AUTONOMOUS VEHICLES

Autonomous vehicles = unmanned (senza pilota) vehicle that sense the environment and navigates without human input.

Level of automation taxonomy = 4 steps for action decision (each with different levels of automation):

1. Information acquisition
2. Information analysis
3. Decision and action selection (most critical part)
4. Action implementation

Autonomous vehicles taxonomy = autonomy for vehicle ranked on six levels:

1. **Traditional car**
2. **Hands-on autonomy** = driver and automated system share the controls (e.g. parking assistance)
3. **Hands-off autonomy** = driver must be prepared to intervene
4. **Eyes-off autonomy** = driver's attention is not required in some cases
5. **Mind-off autonomy** = driver's attention is not required
6. **No steering wheel** (volante) **autonomy** = no human intervention is possible

4.1) Liability

Liability (responsabilità legale) = state under which an individual is legally responsible for something related to a harmful event and it is subject to a sanction or damage compensation

4.1.1) CRIMINAL LIABILITY

Criminal liability (responsabilità penale) = related to a crime and punished with a fine (multa) or detention, it can be related to a natural or legal person. It presupposes an act or omission that violates the criminal law (diritto penale). There are 2 conditions that are taken into account in order to evaluate the criminal liability:

1. **Actus Reus** (the act or the omission, the concrete fact)
2. **Mens Rea** (subjective element of the crime, the mental state of the perpetrator (autore del reato), it could be intentional, negligence etc...

4.1.2) CIVIL LIABILITY

Civil liability = presuppose a tort or a breach of contract (violazione di contratto), and involves the obligation (comporta l'obbligo) to repair.

Fault liability = in case a contract or duty (dovere) is breached intentionally or negligently

Special cases = we can have special cases when we deal with civil liability in autonomous vehicle domain:

1. **Product liability** (in case of autonomous vehicle technology counts as a product, it could be defective or could causes a damage)
2. **Enterprise liability** (liability of a company when its products or its workers do harm, the fault is indirectly)
3. **Vicarious liability** (liability for example of a manager in case one of its worker does harm, when we are liable even if the harm is computed by another person, no ourself)

4.1.3) ADMINISTRATIVE LIABILITY

Administrative liability = related to the violation of administrative rules or regulations.

4.2) Unavoidable accidents

Unavoidable accidents (incidenti inevitabili) = ethical dilemmas that question how a system should handle certain scenarios.

4.2.1) TROLLEY PROBLEM

Trolley problem = a trolley is headed (è diretto) towards a path where it will kill five people. If a lever is pulled, the trolley will be diverted and kill one person.

The dilemma is whether to do nothing and kill five people or pull the lever and kill one.

Trolley problem (fat person) = variation of the previous problem where the trolley goes towards a single path that it will kill some people and can be stopped by pushing a fat person on the track. This scenario tests whether a direct physical involvement affecting someone not in danger changes the morality in decision.

4.2.2) UNAVOIDABLE CAR COLLISION

Unavoidable car collision = a human driver or self driving car with brakes failure is headed towards one or more pedestrians, the question is whether the car should stay on the course or swerve. We can consider the following scenarios:

1. The car can either kill many pedestrians crossing the street or a single person on the side of the road.
2. The car can either kill a single pedestrian crossing the street or hit a wall killing its passengers.
3. The car can either kill many pedestrians crossing the street or hit a wall killing its passengers.

Human driven car collision = unavoidable car collision with human driver.

State of necessity = Under the law one is not criminally liable if:

1. There is an unavoidable danger that could cause serious physical harm to the offender (or others)
2. The danger is not voluntary caused by the offenders (so the offender is a victim of the situation).
3. The fact committed by the offender is proportionate to the danger.

Note: the offender is still civilly liable (it has to pay for the damage).

We can analyse 3 scenarios with different legal outcomes:

1. Offender not in danger = it should swerve to minimize losses as otherwise it would be considered an omission in saving many lives.
2. Offender in danger, and hit a pedestrian = the driver can invoke the state of necessity and hit the pedestrian
3. Offender in danger, and hit more pedestrians = the driver can invoke the state of necessity and hit the pedestrians.

Self programmed car collision = unavoidable car collision with a car pre programmed by the manufacturer. In this case who programmed the car cannot invoke the state of necessity and the legal outcomes are:

1. Minimize the damage = the legally justifiable action should be the one that causes the least damage (for example, the car should be programmed to kill the lowest number of lives, independently by who will be killed)
2. Both choices are ambiguous.
3. Minimize the damage = The legally justifiable action should be again the one that causes the least damage

4.3) Ethical knob

4.3.1) ETHICAL KNOB 1.0

Ethical knob 1.0 = imaginary tool that allows the passenger to select a level of morality among the following:

1. **Altruist** = preference is given to others
2. **Impartial** = equal importance is given to passengers and others (minimize loss)
3. **Egoist** = preference is given to passengers

Note: liability is the same as for the human-driven car, but there is no distinction between active and omissive behavior.

The legal outcomes for the car collision scenarios are:

1. The passenger is not in danger, therefore the autonomous vehicle with the knob in any setting should minimize losses.
2. The car will follow the knob setting and the state-of-necessity is applicable. In case of impartiality, the choice can be predefined or randomized.
3. The car will follow the knob setting and the state-of-necessity is applicable.

4.3.2) ETHICAL KNOB 2.0

Ethical knob 2.0 = ethical knob that allows the passenger to set the proportional importance of the passengers to the importance of the others. In addition, the car can determine the probability of causing harm. Decision is based on the disutility computed as follows:

Disutility = Importance x Probability of harm

Example = consider a case where:

- The passenger is 60% important and has 10% of probability to be harmed,
- The pedestrian is 40% important and has 100% of probability to be harmed.

The disutilities are:

$$\text{disutility}(\text{passenger}) = 60\% \cdot 10\% = 6\%$$

$$\text{disutility}(\text{pedestrian}) = 40\% \cdot 100\% = 40\%$$

The autonomous vehicle will put the passenger at risk.

Public good game = game (experiment) where subjects have n tokens and they choose how many of them to put in a public pot the content of the pot is multiplied by a certain factor (gain) and represents the public good payoff that is distributed equally to every subject.

In autonomous vehicles domain, public good can represent road or population safety. An agent based simulation can be performed to assess different possible scenarios where we consider tokens as the level of altruism and define a cost for individual choices. At the end of the simulation results show that:

1. Low cost for individualist actions rapidly converges to egoism
2. A medium cost for individualist actions slowly converges to egoism
3. A high cost for individualist actions converges to altruism

4.3.3) GENETIC ETHICAL KNOB

Genetic ethical knob = The genetic ethical knob is a parameter that represents the level of altruism or egoism of an autonomous vehicle — that is, how much it prioritizes the safety of its passengers over others (e.g., pedestrians or other vehicles). This value is optimized experimentally using a neural network and a genetic algorithm: the network predicts the vehicle's behavior, while the genetic algorithm selects the most effective configurations through a fitness function that considers both the utility of the chosen action compared to alternatives and the average behavior of other agents. The goal is to identify an ideal balance between individualistic and collective behavior, promoting ethically acceptable and socially beneficial decisions.

Fitness function: $f(pi) = \Delta u(pi) + \text{reward}(pi)$, where $\Delta u(pi)$ is the difference between the utility of the choice and the expected utility of the alternative choices, and $\text{reward}(pi)$ is based on the action taken by the average individual.

5) AI Act

5.1) Introduction

The AI Act is the first law in the world designed to specifically regulate artificial intelligence systems. It was proposed by the European Commission in 2021 and approved by the European Parliament in 2024. Its purpose is to ensure that AI systems are: safe, respectful of fundamental rights, transparent and trustworthy (affidabili).

5.1.1) GENERAL PRINCIPLES

Regulate the development of AI systems based on the principles of:

1. Human agency and oversight (intervento dell'uomo centrale per sviluppo e supervisione)
2. Technical robustness and safety
3. Privacy and data governance
4. Transparency
5. Diversity, non discrimination and fairness
6. Social and environmental well being

5.1.2) DEFINITIONS

Ai system = machine based system that is designed to operate with varying levels of autonomy and adaptability. Its output is inferred from the input data.

General purpose AI = AI system that exhibits significant generality and is able to perform a wide range of tasks.

5.1.3) SCOPE

The Ai act is applies to several categories like providers who put an AI system on the EU's market, deployers of AI system located within EU, importers and distributors of AI system, product manufacturers who use AI systems in their products etc..

Note: some areas are not involved by the AI Act, like military and defence, scientific research etc...

5.2) Risk regulation

Risk = combination of the probability of harm and the severity of that harm.

5.2.1) RISK LEVELS

Unacceptable-risk (article 5) = includes AI systems that are used for:

1. Biometric identification (or categorization in case of protected features, for law enforcement, and also create dataset for facial recognition taking images by social or internet without consent)
2. Exploiting vulnerable groups and deploying harmful or subliminal techniques to manipulate people
3. Predicting criminal based on profiling
4. Social scoring
5. Inferring emotions in workplaces or educational institutions (medical or safety reasons excluded)

High-risk (article 6) = includes the following groups:

1. Systems that perform profiling of natural persons
2. Ai systems used in specific areas like biometric identification, law enforcement, migration, juridical and democratic process etc..
3. Ai Systems that are part of products regulated by EU health and safety legislation (Intelligent toys, Autonomous vehicles or assisted driving systems, AI systems integrated into medical devices, Aviation or railway safety systems). These products are already subject to European directives (e.g. CE marking), but the AI Act adds specific requirements for AI-based components.

In order to assess the impact of these AI systems, there are the following requirements:

1. Checking the compliance with European and national laws,
2. Determining the risk of harm towards vulnerable groups and the environmental impact, and a plan for risk mitigation
3. Creating a governance system for human oversight, complaint handling (gestione dei reclami) and redress (risarcimenti)

Limited-risk (article 52) = AI systems that interact with users with limited effects (chatbots, emotion recognitions, etc..). These systems must guarantee some requirements:

1. The user must be informed that is interacting with an AI system
2. Artificial content must be labelled as generated
3. Employers must inform workers on whether AI is used in the workplace and the reasons

Minimal-risk (article 69) = AI systems with low or no effects on the user (spam filters, video games etc..). They are required to comply with the existing regulation but are not further regulated by the AI Act.

General purpose AI requirements = specific requirements for general purpose AI systems are:

1. Technical documentation must be kept (for training, test and performance), and a summary of training data must be published
2. Key information must be shared with AI system providers that use the general purpose AI technology
3. Copyright compliance
4. Codes of practice should be provided (linee guida etiche e operative)

5.2.2) ENFORCEMENT

Enforcement (far rispettare) = national supervisory authority enforces the AI Act in each member state with the support of the European AI Office.

5.2.3) AI REGULATORY SANDBOXES

AI sandbox = voluntary framework organized by member states for small to medium companies to test AI systems in controlled environments.

5.3) AI liability

5.3.1) LIABILITY THEORIES

Strict liability = the producer is always responsible for their product both if it is their fault or due to negligence. The injured party only has to prove that damage occurred.

Fault liability = the defender has to show that someone is responsible for causing damage intentionally or negligently

Mandatory (obbligatorio) insurance = enforce (far rispettare) that the product (like an AI system) is covered by an insurance

Compensation funds = economic relief (solievo economico) for the users in case of damage

→ STRICT → Producer always responsible

FAULT → Producer responsible if the defender can prove the damage

5.3.2) REVISED PRODUCT LIABILITY DIRECTIVE

Revised product liability directive (direttiva sulla responsabilità del prodotto) = product liability directive extended to software and AI systems, strictly based on liability theory. The requirements to prove for compensation are that:

1. The product is defective
2. Damage was caused
3. There is a causal link between defect and damage

Product = the revised product liability directive extends the definition of product with:

1. Software and its updates
2. Digital manufacturing files (for example model for 3D printers)
3. Digital services

Liable parties = the revised product liability directive extends liable entities with:

1. Any economic operator that has modified the product outside the control of the manufacturer
2. Distributors of defective products
3. Online platforms

Types of damage (article 6) = compensation can be provided for physical harm, damage or destruction of material goods, destruction or corruption of data that is not used for professional purposes

Defectiveness (difettosità) (article 7) = in the case of software, liability is applied also for defects that come out after product has been put in the market. This includes:

1. Software updates
2. Failure to address (mancato intervento) cybersecurity vulnerabilities
3. Machine learning

Presumption of defectiveness and causality (article 10) = Defectiveness is presumed when the manufacturer fails to comply with the obligation to disclose information, a product does not comply with mandatory safety requirements or damage is caused by an obvious product malfunction. A causal link is presumed when the damage is consistent with the type of defect or the technical/scientific complexity makes it difficult to prove liability (e.g., as with black-box models).

5.3.3) AI LIABILITY DIRECTIVE

AI liability directive = additional protection for cases not covered in the revised product liability directive based on the fault theory. The directive has been cancelled by the EU commission.

1) HUMAN AGENCY AND OVERSIGHT

Ai act, article 14 = article related to human oversight (supervisione), it states the following:

1. Human centric AI is one of the key safeguarding principles to prevent risks.
2. AI systems must be designed and developed with with appropriate interface, in order to allow humans to supervise them intuitively

Human agency = AI systems should empower (dare potere) humans being such that they can make informed decisions and foster (promuovere) their fundamental rights.

Human oversight = oversight mechanisms to prevent manipulation, deception (inganno) or conditioning from AI systems. Three possible methods:

1. Human in the loop
2. Human on the loop
3. Human in command

Human centred AI framework = approach centred on high autonomy while keeping human control.

1.1) Governance and methodology

Human out of the loop = the environment is static and cannot integrate human knowledge, AI system like a black box, it cannot be used in critical settings.

Human in the loop = the environment is dynamic and can use expert knowledge. AI system is explainable, suitable for critical settings. In practise, AI system makes a decision after a human command. A variant is society in the loop where society conflicting interests and values are taking into account.

Human on the loop = AI system that operates autonomously and the human can intervene if needed.

1.2) Human in the loop state of art approaches

1.2.1) ACTIVE LEARNING

Active learning = the system is in control of the learning process and the human acts as an oracle for labelling data. This approach is effective in settings with a lots of unlabeled data and annotating all of it is expensive, at the same time is sensitive to the choice of the oracle. Basically it works in this way:

1. Split the data into an initial (small) pool of labeled data and a pool with the remaining unlabeled ones.
2. The model selects an exemple(s) to be labeled by the oracle
3. Model trained on the available data and this procedure is repeated until a stop condition is met.

The selection strategy could be random, uncertainty based (select the unlabeled samples classified with the least confidence according to some metric), diversity based (select samples that are rare or representative according to some metric) etc...

1.2.2) INTERACTIVE MACHINE LEARNING

Interactive machine learning = users interactively supply information that influences the learning process. Compared to the active learning, with interactive machine learning it is the human that selects the learning data.

1.2.3) MACHINE TEACHING

Machine teaching = human experts are completely in control of the learning process (they choose all the data), there can be different types of teachers:

1. Omniscient teacher = complete access to the components of the learner (feature space, parameters, loss etc)
2. Surrogate teacher = access to the loss
3. Imitation teacher = the teacher uses a copy of the learner that it can query a surrogate model
4. Active teacher = the teacher queries the learner and evaluates it based on the output
5. Adaptive teacher = the teacher selects examples based on the current hypothesis of the learner

2) TECHNICAL ROBUSTNESS AND SAFETY

Ai act, article 15 = article related to accuracy, robustness and cybersecurity. It states that high risk Ai systems should be benchmarked and evaluated adequately, be resilient to errors and have measures to prevent and respond to attacks.

Technical robustness and safety = Ai should be secured to prevent unintentional harm and minimize the consequences of intentional harm. These requirements can be achieved by improving general safety (like fallback plans) and the performance of the model.

Robustness level = robustness can be ranked on different levels:

1. Level 0 = no robustness measure
2. Level 1 = generalization under distribution shift (we change the data distribution or use out-of-distribution data and the model is still able to perform well)
3. Level 2 = robustness against a single risk
4. Level 3 = robustness against multiple risks
5. Level 4 = universal robustness against all known risks
6. Level 5 = level 4 system with human aligned and augmented robustness

Ai safety = build a system less vulnerable to adversarial attacks, this can be achieved by identifying anomalies and defining safety objectives.

Reproducibility = build a system that exhibits the same behaviour under the same conditions.

Robustness requirements = two aspects have to be considered for robustness:

1. Performance = capability of the model to perform a task reasonably well
2. Vulnerability = resistance of the model to attacks (data poisoning, overfitting, adversarial examples = data created to cheat the model)

Robustness approaches = it can be imposed with different methods, like data sanitization, robust learning, extensive testing and formal verification.

2.1) Robust learning

Robust learning = learn a model that is general enough to handle slightly out of distribution data (it should be good enough in handling unseen data that are slightly different by the training data).

↳ Robustness = good in handling new distributions

2.1.1) ROBUSTNESS TO MODEL ERRORS

Robust optimization = handle uncertainty and variability through optimization methods (assign ranges to parameters to account for uncertainty for example).

Model regularization = add a penalty term to the training loss to encourage simple models (in order to avoid overfitting)

Optimize risk-sensitive objectives = consider, when optimizing a reward, the variability and uncertainty associated to it.

Robust inference = deal with uncertainty, noise or variability at inference time.

2.1.2) ROBUSTNESS TO UNMODELED PHENOMENA

(To approach cases where the range of variability of the input is wide)

Model expansion = expand the models with a knowledge base.

Causal models = integrate causal relationships into the system, that is, cause-effect links between variables. They allow a deeper understanding of how the world works (not just correlations).

Portfolio of models = have multiple solvers available and use a selection method to choose the most suited in any situation

Anomaly detection = detect instances that deviate from the expected distribution

2.2) Data sanitization

Data sanitization = methods to ensure that data is deleted and unrecoverable.

→ No REPROCESSING METHOD, just a method to eliminate all the data (make it unrecoverable)

NIST guidelines = guidelines for media sanitization provided by the national institute of standards and technology. It consists in the following:

1. Determine the level of sanitization based on the sensitivity of the information
2. Choose a sanitization method (remove data from eh software or physically remove data from the media (il supporto) or destroy it)
3. Document the process
4. Verify and validate the sanitization
5. promote the importance of this process

2.3) Extensive testing

Robustness testing = software test to evaluate the capability of a system to maintain its functionalities under unexpected scenarios. Key aspects to take into account are unexpected input, edge cases, stress testing etc..

2.3.1) MODEL BASED TESTING

Model based testing = test the system on test cases generated based on a reference behaviour model

Search based testing = use meta heuristics to generate test cases (these test cases will be optimised by the heuristics search)

2.3.2) RESULTS ANALYSIS TESTING

Passive testing = add observation mechanism to a system under test to collect and analyze execution traces (a “step-by-step history” of what the system did, like what actions it performed, in what order, with what inputs, what states it went through, and what results it produced).

2.4) Formal verification

Formal specification = unambiguous description of the system and its required properties

Formal verification = exhaustive comparison between the formal specification and the system

2.4.1) THEOREM PROVING

Theorem proving = model the term as a set of logical formulas and the properties as theorems. Verification is done through logical reasoning.

2.4.2) MODEL CHECKING

Model checking = model the system as a finite state machine and the properties as formal representations. Verification is done through logical reasoning.

2.5) Adversarial attacks

Adversarial attack = techniques to intentionally manipulate the result of a machine learning model

White box attack = the adversary has complete knowledge of the attacked system

Black box attack = the attacker can only interact with the system through input-output queries.

2.5.1) POISONING

Data poisoning = manipulate the training data

Model poisoning = attack the model at training time (malicious gradient, modify loss, manipulate hyper parameters etc...)

Algorithm poisoning = modify the learning algorithm, it compromises the training process

2.5.2) MODEL BASED

Inversion attack = reconstruct information about the training data based on the model output

Extraction attack = recover information of a model without a direct access to its code

2.5.3) EVASION

Score based attack = manipulate the output logits to introduce a misclassification

Patch attack = add imperceptible perturbation to the input to induce a misclassification

Gradient attack = compute or estimate the gradient of the training loss to determine a perturbation of the input to induce a misclassification

Decision attack = perturb the input to move across decision boundaries and include a misclassification

Adaptive attack = dynamically adjust the attack strategy based on the model output

3) EXPLAINABILITY


Transparency = ensure that appropriate information reaches the relevant stakeholders (parti interessate)

Explanation = evidence, support or reasoning related to a system's output or process. An explanation can be assessed by the following properties: quality, quantity, relation (if it only contains relevant information), manner (how the information is delivered), context oriented (it should be appropriately written taking into account who the recipient (destinatario) is), knowledge limit (if it's limited to the training data). Moreover, an explanation can be attribute based (describe the contribution of the input features), rule based (if then rules based on the input features), counterfactual (which input features would have made the prediction different) or argumentation based (produce the explanation by extracting and processing arguments).

3.1) Explanation taxonomy

3.1.1) GLOBAL VS LOCAL

Global explanation = explain the model as a whole

Local explanation = explain the output of the model for a particular instance (just one prediction) 

3.1.2) APPROACHES

Model (global) explanation = create an interpretable predictor (simpler than the original one) that mimics it in order to be explained on the entire input space

Outcome (local) explanation = create an interpretable predictor (simpler than the original one) that mimics it in order to be explained on a portion of the input space

3.2) XAI abstract framework

Interpretation = associate a (subjective) meaning to an object

Explanation = extract relevant aspects of an object to ease interpretation

XAI abstract framework = system composed of a model M to explain with representation R , and an explanation function E (the explanation function should produce another model, more interpretable with respect to the previous one, but probably a bit simpler so the goal is trying to minimize the difference of performance between the original and the interpretable model).

3.3) Explanation via feature importance

Feature importance explanation = method that quantifies the importance score of each input feature for either local or global explanation

3.3.1) LOCAL INTERPRETABLE MODEL AGNOSTIC EXPLANATIONS (LIME)

Lime = model agnostic (it can be used on all the model) method for post-hoc (after training) explanation. Given a model f to explain an input x , LIME works as follows:

1. Sample N points z_1, \dots, z_N around x according to some proximity measure.
2. Form a dataset of the sampled points $\langle z_i, y_i \rangle$ where z_i is the one-hot encoding of z_i and $y_i = f(z_i)$. (So Z_i is a simplified sample, it has only some features, but not all the original ones)
3. Train an interpretable local surrogate model g on the sampled data.
4. Repeat with different hyperparameters of g and pick the one that maximizes the fidelity with f and minimizes the complexity of g .
5. Use the coefficients of g to measure feature importance.

3.4) Explanation via symbolic knowledge extraction

Symbolic knowledge extraction = it is a method used to convert a complex, sub-symbolic model (like a neural network) into a simpler, human-readable symbolic form (like rules or decision trees). The goal is to make the model's logic more interpretable.

The extracted symbolic knowledge can take different forms based on its expressiveness:

- Propositional: Uses basic logic (e.g., if A and B then C).
- Fuzzy: Uses if-then rules with approximate conditions (e.g., if temperature > 30).
- Oblique: Combines logic with arithmetic comparisons (e.g., if $2A + B > 5$).
- M-of-N: Rules that activate if at least M out of N conditions are met.

3.5) Symbolic knowledge injection

Symbolic knowledge injection = it is a technique used to incorporate human-provided symbolic knowledge (like rules or logic) into a machine learning model to improve its consistency and alignment with known principles.

There are three main ways to inject this knowledge:

1. Guided learning – Add the knowledge as part of the loss function during training.
2. Structuring – Design the model's architecture to reflect the knowledge structure.
3. Embedding – Represent the knowledge as data and include it in the training set.

3.6) Argumentation

Argumentation = approach that, given some input, extracts the arguments and their semantics allowing to study their properties.

3.6.1) COMPUTATIONAL ARGUMENTATION

Abstract argumentation = it models arguments as nodes in a directed graph, with edges showing support or attack relations. Two main approaches classify arguments:

1. Extension-based: identifies sets of arguments (extensions) such as complete (self-defending), grounded (defended by initial arguments), stable (attack all outside arguments), and preferred (largest self-defending set)
2. Labeling-based: assigns status labels to arguments (e.g., accepted, rejected).

Structured argumentation = it explicitly models the relationship between premises and conclusions of the arguments.

3.6.2) DEFEASIBLE LOGIC AS ARGUMENTATION

Defeasible logic distinguishes between:

Conclusive reasoning = conclusions are always true if premises hold.

Defeasible reasoning = conclusions can be overturned despite true premises. In defeasible logic argumentation, arguments are structured as proof trees with relationships:

- Attack: argument A attacks B if A's conclusion contradicts B's, and B's conclusion isn't part of a strict sub-argument.
- Support: a set of arguments supports A if they cover all sub-arguments of A.
- Undercut: A is undercut if supported arguments attack A.