

Lesson 1_Sartor

Introductory Slides (see Materials)


Castelfranchi gave a general overview of what AI is.

ETHICS GUIDELINES FOR TRUSTWORTHY AI

Document written by the High-Level Expert Group on Artificial Intelligence. This document is a reference for the ethics of AI.

The point of view of this document is a bit different from the idea proposed by Castelfranchi.

Documento che stiamo commentando:

 https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

The document

- Prepared by the High-Level Expert Group on Artificial Intelligence set up by the European Commission in June 2018.
- made public on 8 April 2019.
- available online (<https://ec.europa.eu/digital-single-market/en/highlevel-expert-group-artificial-intelligence>).
- Is a good example of the many documents on ethics of AI published so far → it gives an idea of the stage of the debate
- See also

The idea of trustworthy AI

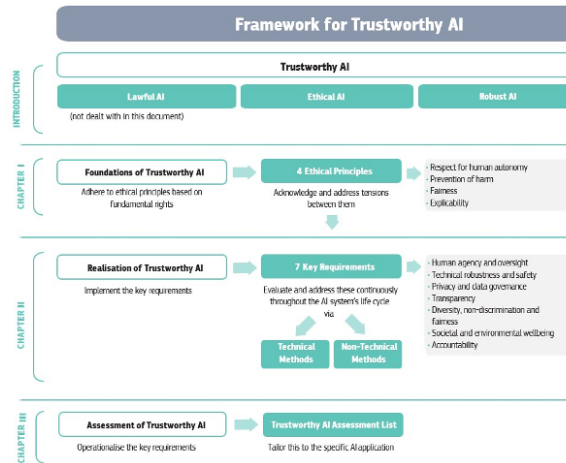
- AI should be
 - Lawful, complying with all applicable laws and regulations
 - Ethical, ensuring adherence to ethical principles and values
 - Robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm
- These requirements should be met throughout the system's entire life cycle

This document offered more a defensive approach, while Castelfranchi was proposing an opportunity (how can we use AI to make our social life better?)

- Can you think of examples of unlawful, unethical or nonrobust uses of AI?
 - In War killing can be a lawful activity, so which kind of AI weapons can be used in war? → In the session of Autonomous Weapons it will be explained
 - If I use a robot to kill somebody, this is an **unlawful** use of the AI
 - Using AI to manipulate people to make them change their political opinion is **unethical** and possibly **unlawful**.

- **Non-robust** use of AI is the case when an AI system makes actions like 'a robot kills a worker in the factory', 'an autonomous car that kill pedestrians and so on' → a system not well developed, in a way that the consequence is to harm people.

→ We have these 3 context: unlawful, unethical, non-robust.



Explanation of the image:

- Introduction: 3 aspects → lawful AI, AI should be ethical, AI should be robust
- Chapter 1: there are the foundations provided by 4 ethical principles → respect for human autonomy, prevention of harm, fairness, explicability

Chapter 1: Ethical principles

- Develop, deploy and use AI systems in a way that adheres to ethical principle:
 1. respect for human autonomy,
 2. prevention of harm,
 3. fairness and
 4. explicability
 - There are 2 ways to understand explicability: one is more computer science directed (= build a model of the system knowing how the system functions and how it answers in a certain way to certain inputs) and one more user oriented (= explaining to the addressee of the decision by an AI system why a certain person is receiving rejection → in this case the system should provide the reason for the rejection. The system succeeds it can say to the person why the decision of the user is rejected and what could do to make the system provide a different output).
- Acknowledge and address the potential tensions between these principles.
 - There are tensions between the 4 principles mentioned above. For example if you want an AI system not to harm, we should limit the way in which people can use it and therefore we limit autonomy → we need to **pay attention to these various principles**
 - Moreover there is usually attention to explicability and performances in AI systems.
- Pay particular attention to
 - **situations involving more vulnerable groups** such as children, persons with disabilities and others that have historically been disadvantaged or are at risk of exclusion, and

- **situations which are characterized by asymmetries of power or information**, such as between employers and workers, or between businesses and consumers.
- Acknowledge that, while bringing substantial benefits to individuals and society,
 - We have to recognize that **AI systems also pose certain risks** (that can and may have a negative impact including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.)
 - Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk.

Chapter II: guidance of realisation trustworthy AI

- Ensure that the development, deployment and use of AI systems meets the seven key requirements for Trustworthy AI:
 - (1) human agency and oversight,
 - (2) technical robustness and safety,
 - (3) privacy and data governance,
 - (4) transparency,
 - (5) diversity, non-discrimination and fairness,
 - (6) environmental and societal well-being and
 - (7) accountability.
- Consider technical and non-technical methods to ensure the implementation of those requirements.

Chapter II: guidance of realization trustworthy AI (continues)

Then the document describes some ideas (some goals) that need to be achieved in order to develop a trustworthy AI.

- **Foster research and innovation**
 - to help assess AI systems and to further the achievement of the requirements; disseminate results and open questions to the wider public, and systematically train a new generation of experts in AI ethics. (this is also the scope of this course —> understand ethical issues involved in the development of AI)
 - In particular in the public, as Castelfranchi said, we should intervene to support AI research that enables ethical goals been achieved. It is important to have a public investment also in how to achieve ethical requirement
- Communicate, in a clear and proactive manner, information to stakeholders about the AI system's capabilities and limitations,
 - enabling realistic expectation setting, and about the manner in which the requirements are implemented. Be transparent about the fact that they are dealing with an AI system.
 - Provide **transparency**: when people use an AI system, they should be aware on that (without being mis-convincing to communicating with a human).
- Facilitate the traceability and auditability of AI systems
 - particularly in critical contexts or situations.
 - This is linked to the idea of **explicability**.
- Involve stakeholders throughout the AI system's life cycle.

- Foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.
- Be mindful that **there might be fundamental tensions between different principles and requirements.**
 - Continuously identify, evaluate, document and communicate these trade-offs and their solutions.

Chapter III: Trustworthy AI assessment

- Adopt a Trustworthy AI assessment list
 - when developing, deploying or using AI systems, and adapt it to the specific use case in which the system is being applied.
- Keep in mind that such an assessment list will never be exhaustive.
 - Ensuring Trustworthy AI is not about ticking boxes, but about continuously identifying and implementing requirements, evaluating solutions, ensuring improved outcomes throughout the AI system's lifecycle, and involving stakeholders in this.

The Commission's approach to AI

- Communications 25 April 2018 and 7 December 2018 (COM(2018)237 and COM(2018)795). Three pillars:
 - (i) increasing public and private investments in AI to boost its uptake
 - (ii) preparing for socio-economic changes, and
 - (iii) ensuring an appropriate ethical and legal framework to strengthen European values.
- <https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM2018237F1ENMAINPARTPDF>
- <https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM2018795F1ENMAINPARTPDF>

Ethics is just an aspect: in Europe we do not want only to be ethical but we want also be able to play a leading role in development of AI → is important to increase private and public investment in Ai and prepare economical changes.

Issue: are we really able to match US and China?



<https://ec.europa.eu/growth/tools-databases/dem/monitor/content/usachina-eu-plans-ai-where-do-we-stand>

Now Europe is not at the same level of investment/development as North America or Asia.

Human-centric AI

- commitment to the use of AI in the service of humanity and the common good, with the goal of improving human welfare and freedom.

- Maximise the benefits of AI systems while at the same time preventing and minimising their risks.

These 2 points are obvious in theory but not in practice since the development of AI is upon large companies who aim at increasing the profit and control markets and profits —> so in practical these 2 principles are not so obvious.

Ethics vs law

What is ethics and how is different from general decision making?

In our decisions we can have our particular interests —> we engage in ethical reasoning when we try to determine what should be done having some kind of **impartial consideration of all interests at stake**.

- **Ethics:** norms indicating what should be done, with regard to all interests at stake
 - **Positive ethics:** norms shared in a society (possibly including ideas of social hierarchy, gender roles, etc.)
 - There can be many norms in a society and these kind of norms can be good or bad basing on a perspective.
 - **What people believe that is good**
 - **Critical ethics:** norms that are viewed as most appropriate, or rational
 - For example feminists would have argued some years ago that the idea that women should be subordinated to men (this was an ethical rule shared among the most of the people). A critical attitude could be to go against such rule and say that is not good because is not appropriate.
 - **What people criticize to consider a norm bad, unfair etc.**
- **Law:** norms that adopted through institutional processes and coercively enforced.

Ethics may be more demanding than law: for example in the instance of advertising there are certain behaviours that are legally permitted but they are unethical.

What is the role of ethics given the difference of speed between law and technology?

The Guidelines for Trustworthy AI as a (critical) ethics?

- The guidelines are addressed to all AI stakeholders designing, developing, deploying, implementing, using or being affected by AI,
 - including but not limited to companies, organisations, researchers, public services, government agencies, institutions, civil society organisations, individuals, workers and consumers.
- "Nothing in this document shall create legal rights nor impose legal obligations towards third parties. We however recall that it is the duty of any natural or legal person to comply with laws – whether applicable today or adopted in the future according to the development of AI."
- What is the role of ethics, relatively to law in the AI domain?

AI should be lawful

- It should comply with
 - EU primary law (the Treaties of the European Union and its Charter of Fundamental Rights),
 - EU secondary law (regulations and directives, such as the General Data Protection Regulation, the Product Liability Directive, the Regulation on the Free Flow of NonPersonal Data, anti-discrimination Directives, consumer law and Safety and Health at Work Directives),
 - UN Human Rights treaties and the Council of Europe conventions (such as the European Convention on Human Rights)

- Laws of EU Member State laws (Italian law).
- Laws can be horizontal or domain-specific rules (e.g., on medical devices)
- Issue: Can you think of a horizontal law covering all AI applications?

Foundations of trustworthy AI

- **AI ethics is a sub-field of applied ethics,**
 - **focusing on the ethical issues raised by the development, deployment and use of AI.**
 - **Its central concern is to identify how AI can advance or raise concerns to the good life of individuals, whether in terms of quality of life, or human autonomy and freedom necessary for a democratic society.**

Foundation: (Ethical) fundamental rights

In this document there is a discussion on the link between **ethics and fundamental rights**.

The reference here is to basic fundamental right (such as human dignity etc → the one mentioned in the list below)

Fundamental right: these are the human rights that every person has as a human being according to international law.

There are also the fundamental rights that every person belonging to a certain political community has, with regard to the state of their community.

Here the main reference is the chapter of the Fundamental rights of the European Union. Human rights (or fundamental rights as well) have a double role to play: they are established by the law but they are argued also for based on ethical importance.

- **Respect for human dignity** (human dignity = the value that a person is assumed to be). Human dignity encompasses the idea that every human being possesses an “intrinsic worth”.
 - The idea is that every human being has some importance and when dealing with a human person. For example you may wonder that if AI is cheating a person into doing something that is known to be bad for that person for the profit purpose, in this case you may say that the dignity of this person is not adequately taken into account because the interest of this person is not all the way in which the person is treated → this person is just the mean for a goal that goes against his/her interest.
 - This idea is connected to the philosophy of Kant.
- **Freedom of the individual.** (Related to the concept of autonomy) Human beings should remain free to make life decisions for themselves: including (among other rights) protection of the freedom to conduct a business (is also a fundamental right according to chapter of the European Union, probably not based on the convention of Universal Declaration of Human Rights, because at that time communism was a powerful ideology), the freedom of the arts and science, freedom of expression, the right to private life and privacy, and freedom of assembly and association.
 - We will examine in depth how AI may impact on this important right.

- This idea is connected to the philosophy of Kant.
- **Respect for democracy, justice and the rule of law.** AI systems must not undermine democratic processes, human deliberation or democratic voting systems, due process and equality before the law
 - human deliberation or democratic voting system: An example is the case in which it was used AI to predict peoples' political attitude and then use this prediction to modify people's opinion —> this goes against the principles of dignity and autonomy
 - due process and equality before the law: use of AI systems in judicial systems. Use AI to make predictions concerning people likelihood to commit a crime and consequently determining how much time they have to stay in prison.
- **Equality, non-discrimination and solidarity** - including the rights of persons at risk of exclusion. In an AI context, equality entails that the system's operations cannot generate unfairly biased outputs. (GS: we need to understand what this means)
 - Need to take care that people of a certain group is not rejected from an AI system when apply for a certain position or for a credit. We have to consider what means 'being biased in a computer system or in whatever system making predictions'
- **Other citizens' rights:** the right to vote, the right to good administration or access to public documents, and the right to petition the administration

Ethical principles (based on human rights)

These are the fundamental ethical principles mentioned in this document and they refer to bioethics.

- **Bioethics** is the ethics concerning the medical domain (such the discussion concerning eutanasia or medical intervention on human body and human genome).
- (i) Respect for human autonomy
- (ii) Prevention of harm
- (iii) Fairness
- (iv) Explicability

(i) Respect for human autonomy

- Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process.
 - **AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans.**
 - Examples of kind of systems that are unjustifiably subordinate, coerce etc: use of face recognition to identify people in public spaces and subjects to do surveillance. In some some countries the use of face recognition is non-ethical or unacceptable while for example in China is accepted.
 - **they should be designed to augment, complement and empower human cognitive, social and cultural skills.**
 - Here there is the issue that when we rely on AI we can become incapable of exercising our cognitive skills. As today we no longer need to know the streets when going around because we have our GPS, we can imagine that our AI might govern our life to delegate to AI systems all those kind of cognitive activities and we leave the cognitive activity to manage ourselves. This is the **fear** that was also the motivation of the 'Robot Saga'.
 - **The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice.**

- this concerns the use of AI in the work domain where humans should have the work improved with the coordination with AI rather than be subject to control of an AI system. This is the case of work managed by platforms, which is the case of uber drivers or food delivery —> they are controlled by computer systems that determine their trajectories and assign them rewards or punishment in case they don't behave as expected. These of micro-control can be heavy for humans. The same happens in the Amazon Storage centers.
- **This means securing human oversight over work processes in AI systems, supporting humans in the working environment, and aiming for the creation of meaningful work.**
 - This is a very important requirement that we have to implement in the future

(ii) The principle of prevention of harm

- **AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings.**
 - This entails the protection of human dignity as well as mental and physical integrity.
 - AI systems and the environments in which they operate must be safe and secure. (avoid accidents of robots industries and so on)

An example is the use of AI in fabrics, in which there were cases of humans killed by the robots. An other example is the case in which people are killed by autonomous cars.

Is impossible to ensure full safety, but the use of AI should provide an improvement and not a risk for safety of people (so not a reduction of safety for the people).

(iii) The principle of fairness

This is a very complex-key issue.

We don't have a real term to full capture the idea of fairness.

We decompose the dimensions of **fairness**.

- **Substantive dimension**
 - ensuring equal and just distribution of both benefits and costs, and
 - ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.
 - Promoting equal opportunity in terms of access to education, goods, services and technology.
 - AI can contribute to that for example making cheaper and more flexible social services.
 - Never leading to people being deceived or unjustifiably impaired in their freedom of choice.
 - Here we may consider the use of AI systems to deliver aggressive or misleading advertisements when we are online. Consider for example the case in which we want to buy an online ticket and we receive a message that says 'this is the last ticket, 'this is the best price'.
 - AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives
 - when you are developing a system you have to consider the costs involved both in the development of the system and the costs to reach interest and objectives.
 - For example have face recognition in the streets can improve security, but we have to consider that with face recognition everybody is known when they move around and this opens for abuses.
- **Procedural dimension.**

- ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them
 - In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

We should identify who is accountable for the decision (a collective/a company), the people who are responsible and a person that when the system is alone explains a certain behaviour of the system. —> explAinability is one of the principles of bioethics. **Key issue in AI systems.** Approaches based on deep learning today deliver decisions that are appropriate/efficient in many cases but it's difficult to provide an adequate explanation.

We need a tradeoff between efficiency of the systems and explicability of the outputs.

In certain domains explicability is more important than in others.

(iv) The principle of explicability

Explicability is important in decisions concern humans because we need to ensure **contestability**: is someone is unhappy of the decision he has received, the he should be able to get an answer. Before the answer it should be known how the system work and after the decision i should get an explanation of why this decision has been taken in this particular way.

- To ensure **contestability**
 - processes need to be transparent,
 - the capabilities and purpose of AI systems openly communicated, and
 - decisions – to the extent possible – explainable to those directly and indirectly affected.
 - —> being able to predict the output decision and explain the output decision.
- **An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible.**
 - other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights.
 - thee degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.

Tensions between the principles

- Methods of accountable deliberation to deal with such tensions should be established.
 - There is a tension between **explicability and performance** in some domains.
 - Conflicts between **prevention of harm and human autonomy**
 - for example more surveillance may allow to catch more easily the criminals, but it would affect the freedom of actions of the people being supervised.
 - Also between **welfare and security**?
 - for example put accessing limitation on products or on the way in which AI works may make them less productive.

Requirements of Trustworthy AI

1. Human agency and oversight

- Including fundamental rights, human agency and human oversight

- It goes behind the idea of autonomy

2. Technical robustness and safety

- Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
 - This is important in relation to some of the values previously mentioned such as prevention of harm which includes also all the things mentioned in the previous point (resilience to attack etc)

3. Privacy and data governance

- Including respect for privacy, quality and integrity of data, and access to data
 - **Privacy:** enabling people to have some control over personal data
 - In Europe we have GDPR for regulation of privacy.

4. Transparency

- Including traceability, explainability and communication
 - Ability to know how the system works and have an explanation of its outcomes.

5. Diversity, non-discrimination and fairness

- Including the avoidance of unfair **bias**, accessibility and universal design, and stakeholder participation
 - issue of **biases**: it happens when a particular group is negatively affected in a particular way in the functioning for which the AI system is developed
 - AI systems for face recognition example: it happens quite often that black peoples are accused to be criminal due to low accuracy of systems affected by biases. False positives in the search for criminals.
 - Article about the issue of **face recognition** (the researcher that wrote the article was dismissed by Google for expressing certain opinions on the matter fo the AI and ethics):

A Second AI Researcher Says She Was Fired by Google

For the second time in three months, a prominent researcher on ethics in artificial intelligence says Google fired her. On Friday, researcher Margaret Mitchell said she had been fired from the company's AI lab, Google Brain,

<https://www.wired.com/story/second-ai-researcher-says-fired-google/>



Study finds gender and skin-type bias in commercial artificial-intelligence systems

Three commercially released facial-analysis programs from major technology companies demonstrate both skin-type and gender biases, according to a new paper researchers from MIT and Stanford University will present later this month at the Conference on Fairness,

<https://news.mit.edu/2018/study-finds-gender-skin-type-bias-artificial-intelligence-systems-0212>



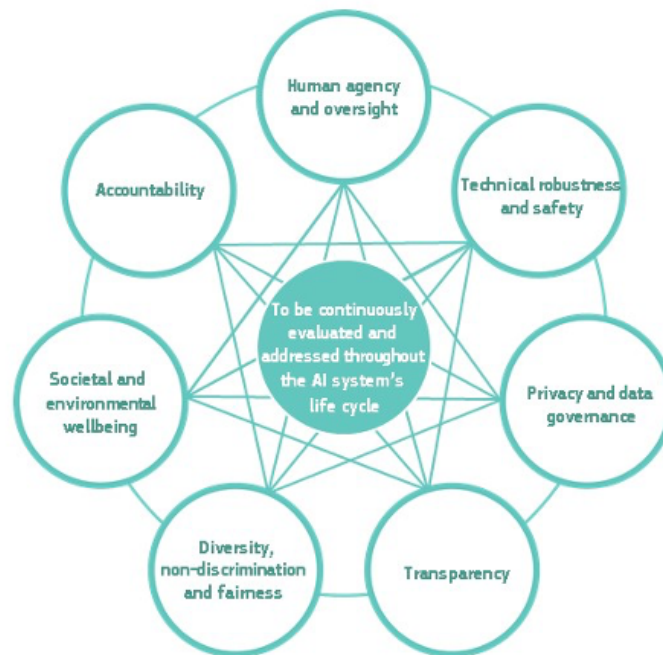
6. Societal and environmental wellbeing

- Including sustainability and environmental friendliness, social impact, society and democracy
 - Important issue: environmental impact of Machine Learning. Some ML models require a vast amount of computation and computation impact also the use of energy and has therefore a negative impact on the environment, especially when this energy is produced by fossile or gas.
 - There is also an environmental positive impact of AI on environment when AI enables a better use of energy for example by the use of smart-grids.

- negative impact if it AI is used for manipulation in political election, but there are positive uses when AI facilitates in detecting unlawful practices or by facilitating communication, interaction, decision making.

7. Accountability

- Including auditability, minimisation and reporting of negative impact, tradeoffs and redress.
- There are various mechanism for accountability and the idea is that **also AI should be accountable**.
 - there is accountability when if something goes wrong there is someone you can ask the reason and in case this explanation is insufficient, this person can brought in front of a forum that can express a judgement on the accountable person.
- There can be also a **legal accountability**: if something is done and is considered illegal, then again there are legal institutions and judging tools where people provide explanations and in case the explanation is not satisfactory, they may find a legal responsible.



Human agency and oversight

- AI systems should support (but **not substitute**) human autonomy and decision-making
 - this a a problem which is present whenever AI systems substitute people —> as in aviation systems. —> Problem that happens in the domain of **aviation-automation**: systems that automate flight control and pilot loose the activity to fly the airplane. Idea should be, rather than substitute humans, to create a support so that humans can remain in control. This could be good for the performance of the system which integrate advantages of system and human capability.
- So when an AI system is introduced it should be studied its impact on the society (Human-rights assessment, human agency, human oversight, technical robustness and safety etc)

Therefore they should support

- **Fundamental rights**
 - Human rights assessment Human agency.
- **Human agency**
 - Users should be able to make informed autonomous decisions regarding AI systems.
- **Human oversight.**
 - Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects (human-in-the-loop (**HITL**), human-on-the-loop (**HOTL**), or human-in-command (**HIC**) approach + public controls)
- **Technical robustness and safety**
 - AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm.
- **Resilience to attack and security**
 - AI systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries
- **Fallback plan and general safety**
 - AI systems should have safeguards that enable a fallback plan in case of problems
- **Accuracy**
 - AI systems should have the ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models.
- **Reliability and Reproducibility .**
 - The results of AI systems should be reproducible (given the same inputs, the AI system should produce the same outputs), as well as reliable.

Privacy and data governance

- **Prevention of harm necessitates privacy and data governance:**
 - Privacy and data protection.
 - AI systems must guarantee privacy and data protection throughout a system's entire lifecycle.
- **Quality and integrity of data**
 - The data used to train a systems should not contain socially constructed biases, inaccuracies, errors and mistakes, malicious data should not be added
- **Access to data**
 - Data protocols governing data access should be put in place.

Transparency

- This requirement is closely linked with the principle of explicability (GDPR article 22)
- **Traceability**
 - The data sets and the processes that yield the AI system's decision, should be documented
- **Explainability**

- The technical processes of an AI system and the related human decisions should be explainable
- **Communication**
 - Humans have the right to be informed that they are interacting with an AI system.

Diversity, non-discrimination and fairness

- We must enable inclusion and diversity (very debated nowadays) throughout the entire AI system's life cycle
 - **Avoidance of unfair bias**
 - Prevent unintended (in)direct prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation, due to data or algorithms
 - There are various theories : there is also a debate among compute sciences about what fairness is a computer
 - **Accessibility and universal design**
 - AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics
 - **Stakeholder Participation**
 - Open discussion and the involvement of social partners and stakeholders, including the general public
 - **Diversity and inclusive design teams**
 - the teams that design, develop, test and maintain, deploy and procure these systems reflect the diversity of users and of society in general

Societal and environmental well-being

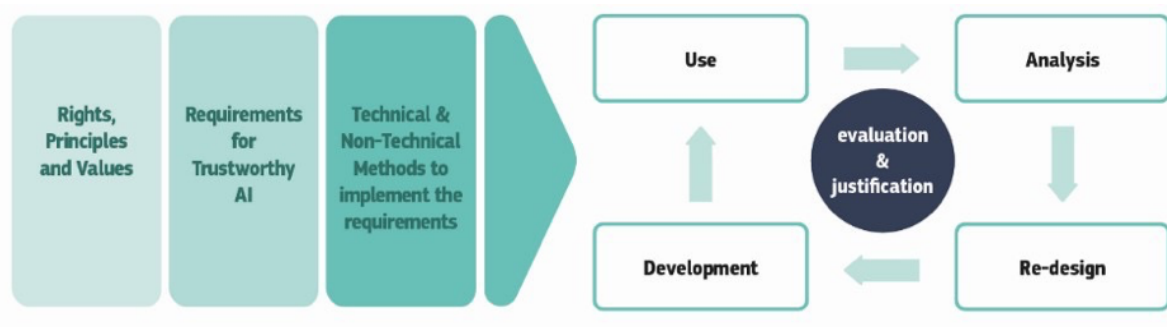
AI impacts many aspects of our society so a very important issue is the environmental friendliness of AI

- The broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle.
 - **Sustainable and environmentally friendly AI**
 - Measures securing the environmental friendliness of AI systems' entire supply chain should be encouraged.
 - Developing models used for Transfer Learning regard a huge amount of energy and we need to move to a more sustainable AI
 - **Social impact.**
 - The effects of these systems on individuals, groups and society must therefore be carefully monitored and considered.
 - Impact of AI in political propoganda and impact of AI in determining what kind of news or information each person accesses from Facebook or Google —> use of AI to engage people is used to address contents to people dividing people in big bubbles completely separated. —> AI contributes to split society into non-communicating silos. AI contributes in the issue of certain bubbles because AIs decide which contribution of information provide to each person.
 - **Society and Democracy.**
 - Take into account AI's effect on institutions, democracy and society at large

Accountability

- Ensure responsibility and accountability for AI systems and their outcomes
 - **Auditability**
 - It should be possible to enable the assessment of algorithms, data and design processes
 - Attention between the secrecy (industrial secrecy) to not lead intellectual property and enable people to assess how the system works. Problem of "Open source vs proprietary software " comes also in AI
 - **Minimisation and reporting of negative impacts**
 - The ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured.
 - If something goes wrong we need to communicate to someone
 - **Trade-offs**
 - Trade-offs should be addressed in a rational and methodological manner within the state of the art
 - We already mentioned that there is the problem of finding tradeoffs between different conflicting requirements. We need to find for example a tradeoff between privacy and security, explainability and efficiency and so on. Unfortunately we cannot have all of everything and so a balance may be done. The tradeoff is unavoidable but it needs to be clear and done in a reasonable way.
 - **Redress**
 - Accessible mechanisms should be foreseen that ensure adequate redress

Technical and non-technical methods to realize Trustworthy AI



Questions and suggestions

- Questions
 - Has the Trustworthy AI document provided you with useful indications?
 - Do you think that they are concretely applicable?
 - Are ethical guidelines that are not legally binding really useful?
 - Any specific criticism?
- Suggestions
 - Read all the document!
 - Read also

Generally, most of the things written in this document respect the classical legal thinks (as we can't kill someone also and AI system cannot kill someone) —> legally binded

But other things should also not be legally binded

Non legally binded requirements have a chance to be voluntarily implemented by AI experts or no?

I have a "tricky" question. These ethical principles seem to be related to our Western culture, however what about other cultures, whose ethical principles might be orthogonal or opposite to ours? In Asia for example can be different the balance between the concepts but not the concepts themselves.

Reply of the professor: I think that these requirements should be quite universal, the problem is that different cultures generally recognize the same needs (for example harming people would be bad for everybody, but a different use of how social interests would override the respect for human autonomy. We have an example concerning the virus: to what extent is ok to put a draconian measure limits human autonomy to prevent for the purpose of help?

Also fairness is a general requirement and can be differently understood.

So in conclusion we have discussed the issue of relativism vs universality with these requirements, but first of all this ethical chapter is directed to European researcher. In Asia some principles could be different, but what is different in general, more than the principle is the way to balance, so the importance that one principle has in relation to another.

Another topic to discuss is that **ethical documents are sometimes criticized** and is said Google, Amazon etc like ethical documents because ethical documents are an alternative to the law. What these companies dislike is coercive measures being imposed by the law, while ethical documents leave them with vaster degree of freedom.

So there are different opinions: for some people ethical documents are only like window dressing, so that the current practice can go on as it is, while for others ethical documents are useful components to make legally binded documents and they maybe show possible ways in which the law could also be developed —> we have this kind of optimistic approach in the second group of observers.