Orlandi Mattia

ETHICS IN ARTIFICIAL INTELLIGENCE

→1 SCIENCE-ORIENTED AI:

For a science-oriented AI not servant of the business (talk):

We'll live in an "augmented" world, in which the real and the virtual are mixed, and in a hybrid society, in which artificial intelligences (not only robots but intelligent software agents too) acts as our cognitive prostheses.

AI is not just building a new technology but a new socio-cognitive-technical system (anthropological revolution). As social engineers, we must focus on two aspects of AI:
1. the importance of the "science" of AI;
2. problems and dangers of the digital revolution, of the "mixed" reality and "hybrid" society.

① For a science-oriented AI

- In general, the main focus of AI research should primarily be Knowledge, not just applications and technology. However, AI still has a too strong technical identity rather than a science identity.
- AI provides conceptual and cognitive instruments for modelling and understanding minds, intelligence and interactions.
- The economic, social and technical outcomes should mainly be side effects of AI research, not the main goal.
- The scientific advantages of the artificial, synthetic approach to mind and society is understanding by building and simulating.
- Aims of "scientific AI":
  - model and explain human and natural intelligence;
  - emulate them;
  - create new intelligence and its theory (General Intelligence).
- Philosophers frequently claim that what AI and cognitive scientists are doing is "anthropomorphize" machines by simulating natural intelligence, whereas it's actually the other way around: AI tries to "de-anthropomorphize" such concepts by making them no longer anthropocentric but more general, abstract and formalized.
- AI mission is not to just borrow concepts and theories from human and social sciences or philosophy and to apply them to technology, but it must change such concepts, models and theories.
- Our brain and minds will be augmented (evolution of social cognition):
  - collective intelligence and problem-solving;
  - collective sense-making (i.e. our interpretation of events);
  - collective knowledge capital and sharing;
  - collective creativity;
  - a new "embodiment" of our cognitive representations;
  - externalized and distributed cognition and mind
- One of the main functions of the brain is integrating and augmenting the perceived reality with memories and expectations.

② Who the AI revolution is empowering

- We are responsible for the introduction of agents, which:
  - are autonomous (proactive) and social;
  - cooperate with humans by following norms (but also violating them if strictly necessary);
  - critically adopt our goals (not just execute orders, but "over-help").
- We must be aware of possible appropriation and unacceptable uses of these instruments.

E.g. Autonomous weapons that can make killing decisions on their own.

- We should implement moral agents, with internalized ethical values to guide their actions.

2.1 Business-oriented AI

- Recently, AI has always been associated to industry and business, whereas it should deal with also moral and political philosophy and social sciences.
- If AI is beneficial for profit and business interests, it's not necessarily beneficial for workers.
- AI can be very beneficial for:
  - democracy;
  - good market with reduced deception and manipulation;
  - social planning;
  - transparency.

2.2 Hidden interests and awareness

- Security, privacy, war and ethics are sure very relevant issues, but not the most or the only relevant ones from the moral and political point of view.
- Hidden interests, manipulation of users/programmers, and in general emptying democracy are not less important, and scientists have to be aware of such aspects.
- Democracy is not a formal and misinformed voting ritual, there is the need of raising collective awareness and encouraging rational decision making (in which we ask ourselves in favor of whom we're acting).
- Intelligent agents must help us understand not only our goals and how to rationally decide, but also who we're favouring.
- Moreover, the goals of intelligent agents should be as transparent to us as possible: they must be able to explain us the reasons behind their decisions ⇒ cognitive model of "reasons" for goal processing and decision making.
- In a lot of circumstances AI will either:
  - decide for us;
  - give us recommendations or a "little push";

  but they won't play a tutelary role, taking care of our good even if in conflict with us.

E.g. Recommender systems are just personalized advertising and act more in favour of the seller than of the user.

- Tutelary means caring of our individual personal interests, and also helping us understand:
  - common interests and collective subjects;
  - hidden conflicts of interests;
  - public good.
- Augmented intelligence also means augmented social awareness

③ Mouth of truth

- We're developing algorithms for ascertaining the "truth" in all the data that's available online ⇒ problem of deciding on which base such algorithm considers a source as reliable.
- The algorithm should be able to distinguish between a conflict of values/interests and a mere conflict between more or less credible data

④ Presences in the mixed reality

- The autonomous and proactive intelligent entities will become presences and roles in the hybrid society and mixed reality.
- This raises the problem of how to manage these autonomous and too informed intelligent agents:
  - which roles such entities will play in our life and environment (tutors, supervisors or "tempting spirits");
  - whether we'll incorporate them as our mental prostheses, listening to their "voice" as our own mental voice (expanded super-ego), or they'll be externalized.

"reflexively social" solutions: augmented internalized self and consciousness.

"social" solution: externalized voices and agents.

⑤ Disagreement technology

- There's a too strong ideology and rethoric about society as cooperation and common intents, and the web has favoured a deviating political feeling of "we against them".
- However, there is no "we" with common values and goals which has to be unified against the political power.
- Population is composed of different classes, genders, generations and cultures with very different and conflicting values and interests.
- Political forces were supposed to represent and protect those different interests, not just the "common" interest: some conflicts of interests can be solved and reconciled in a common interest, but a large part of political/government decision is not for a common advantage.
- Conflicts are not just of views or opinions, but of objective interests too: social conflicts in fact do not have a verbal, cognitive or technical solution based on data and technical principles, they have a political solution based on compromises and equilibrium.
- Conflicts are necessary for democracy and progress, as they can change society in favour of disadvantaged classes.
- In democracies there's the tendency to vote in a self-defending way, and political education is not enough.
- One of the main tasks of AI social technologies should be making conflicts emerge s.t. people become aware of them.
- Using web technologies to organize movements is not so good without promoting critical thinking and counteracting confirmation bias, prejudices and the "bubble effect".
- AI should be used to encourage critical thinking.
- Net interaction is perceived as non-hierarchical, without a superstructure, spontaneous and thus "free", "democratic".
- However, such perception is wrong, as data on the web are often exploited to manipulate users.
- There is the need of anti-manipulation AI technologies: a tutor inducing me to understand and to reflect about why I am oriented in a certain direction, making me conscious, instead of an entity persuading me towards certain choices.

⑥ Concluding remarks

- The revolution of ICT, of digital monitoring and predicting (by simulation), and of big data, can give to society a "glass where to observe themselves", a "glass of the invisible" reflecting also hidden prescences and future predictions.
- AI could help raising our awareness and "making the invisible, visible".

Ethics guidelines for trustworthy AI:

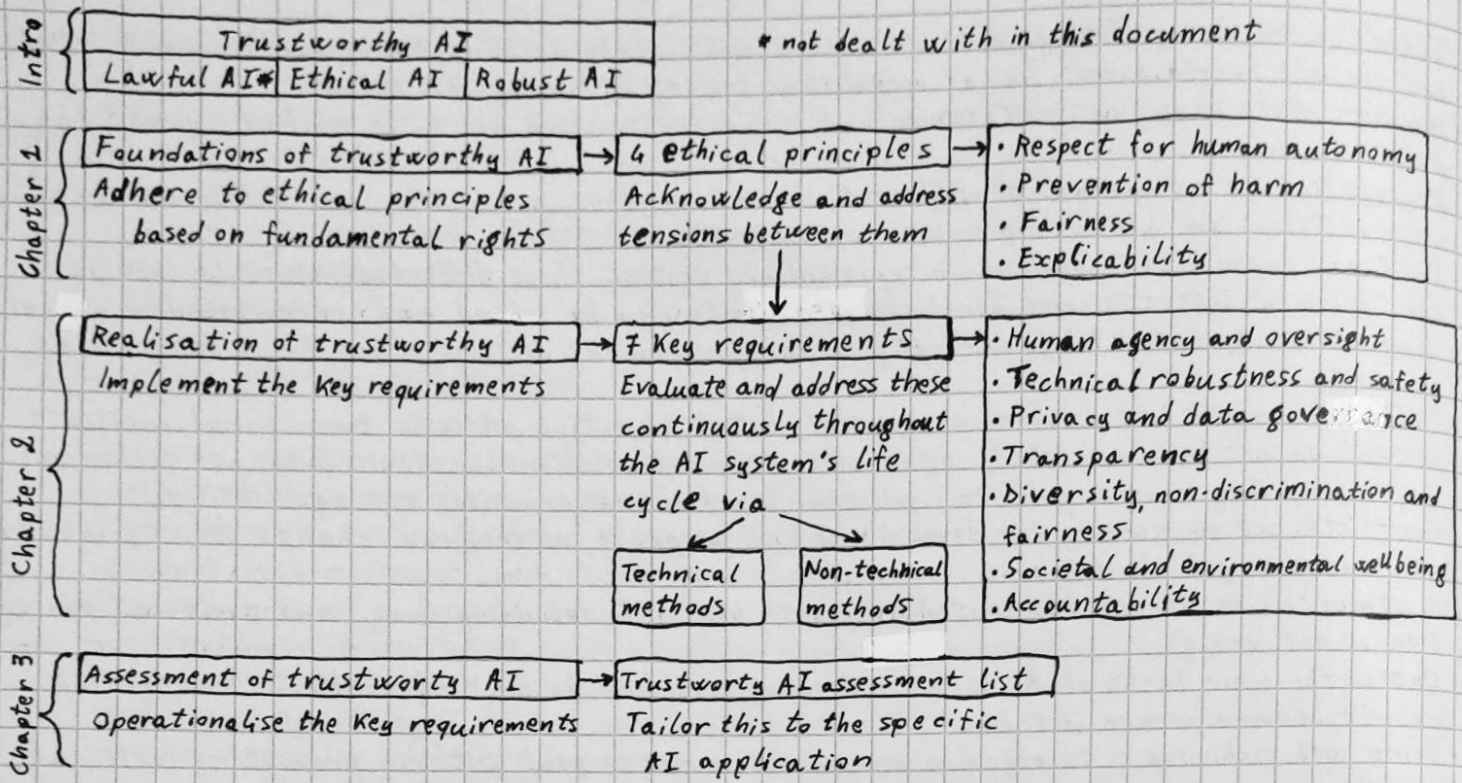
This document was prepared by the High-Level Expert Group on Artificial Intelligence set up by the European Commission in June 2018.

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.

E.g. An AI system that performs cyber attacks in unlawful, one that manipulates people's political opinions is lawful but unethical; an autonomous car that hits a pedestrian is not robust.

Framework for trustworthy AI:



① Chapter 1: Ethical principles

- Develop, deploy and use AI systems in a way that adheres to ethical principles:
  - respect for human autonomy;
  - prevention of harm;
  - fairness;
  - explicability.
- Acknowledge and address the potential tensions between these principles.
- 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;
  - situations which are characterised 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:
  - AI systems also pose certain risks 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);
  - there's the need to adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk.

② Chapter 2: guidance of realisation of trustworthy AI:

- Ensure that the development, deployment and use of AI systems meet the seven key requirements of trustworthy AI:
  1. human agency and oversight;
  2. technical robustness and safety;
  3. privacy and data governance;
  4. transparency;
  5. diversity, non-discrimination and fairness;
  6. environmental and societal well-being;
  7. accountability.
- Consider technical and non-technical methods to ensure the implementation of those requirements.
- 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 [redacted] ethics.
- 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 (also, be transparent about the fact that they are dealing with an AI system).
- Facilitate the traceability and auditability of AI systems, particularly in critical contexts.
- Involve stakeholders throughout the AI system's life cycle, and foster training and education s.t. all stakeholders are aware of and trained in trustworthy AI.
- Be mindful that there might be fundamental tensions between different principles and requirements, and thus continuously identify, evaluate, document and communicate these trade-offs and their solutions.

③ Chapter 3: 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, and that 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, while involving stakeholders in this.

④ The Commission's approach to AI

- Three pillars:
  - increasing public and private investments in AI to boost its uptake;
  - preparing for socio-economic changes;
  - ensuring an appropriate ethical and legal framework to strengthen European values.
- Nowadays, Europe is behind US and Asia w.r.t. investments on AI research.

⑤ 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.

⑥ Ethics vs Law

- 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.)
  - ↳ critical ethics ⇒ norms that are viewed as most appropriate or rational
- Law ⇒ norms that are adopted through institutional processes and coercively enforced

⑦ Guidelines for trustworthy AI and ethics

- Stakeholders committed towards achieving trustworthy AI can voluntarily opt to use these guidelines as a method to operationalise their commitment.
- 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".

⑧ AI should be lawful

- AI 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 Non-Personal 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);
  - EU Member State laws (such as the Italian law).
- Laws can be horizontal or domain-specific rules (e.g. on medical devices).

⑨ Foundations of trustworthy AI

- AI ethics is a sub-field of applied ethics:
  - it focuses 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.
- Ethical fundamental rights:
  - respect for human dignity ⇒ human dignity encompasses the idea that every human being possesses an "intrinsic worth"
  - freedom of the individual ⇒ human beings should remain free to make life decisions for themselves, including (among other rights) protection of the freedom to conduct a business, the freedom of the arts and science, freedom of expression, the right to private life and privacy, and freedom of assembly and association
  - 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
  - equality, non-discrimination and solidarity, including the rights of persons at risk of execution ⇒ in an AI context, equality entails that the system's operations cannot generate unfairly biased outputs
  - other citizens' right to vote, the right to good administration or access to public documents, and the right to petition the administration

⑩ Ethical principles based on human rights

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. Explicability

10.1 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;
- they should be designed to augment, complement and empower human cognitive, social and cultural skills;
- the allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice;
- 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.

10.2 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.

10.3 Fairness

Substantive dimension:
- ensuring equal and just distribution of both benefits and costs;
- 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;
- never leading to people being deceived or unjustifiably impaired in their freedom of choice;
- AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives.

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.

10.4 Explicability

To ensure contestability:
- processes need to be transparent;
- the capabilities and purpose of AI systems openly communicated;
- decisions, to the extent possible, explainable to those directly and indirectly affected.

However, 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;
- the 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.

Obs. there are tensions between the principles (e.g. automated surveillance system can prevent harm but also undermine human autonomy).

(11) Requirements of trustworthy AI

1. Human agency and oversight (including fundamental rights)
2. Technical robustness and safety (including resilience to attack and security, fall-back plan and general safety, accuracy, reliability and reproducibility)
3. Privacy and data governance (including respect for privacy, quality and integrity of data, and access to data)
4. Transparency (including traceability, explainability and communication)
5. Diversity, non-discrimination and fairness (including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation)
6. Societal and environmental well-being (including sustainability and environmental friendliness, social impact, society and democracy)
7. Accountability (including auditability, minimisation and reporting of negative impact, trade-offs and redress)

11.1 Human agency and oversight

AI systems should support human autonomy and decision-making, therefore they should support
- fundamental rights ⇒ human rights assessment;
- human agency ⇒ users should be able to make informed autonomous decisions regarding AI systems;
- human oversight ⇒ it helps ensuring that an AI system does not undermine human autonom or causes other adverse effects;
  - ↳ human-in-the-loop (HITL)
  - ↳ human-on-the-loop (HOTL)
  - ↳ human-in-command (HIC) + public controls
- technical robustness and safety ⇒ AI systems should be developed with a preventative appr ach to risks and in a manner s.t. they reliably behave as intended while minimising unintentional and unexpecte harm, and preventing unacceptable harm.

11.2 Technical robustness and safety

AI systems should match the following criteria:
- resilience to attack and security ⇒ they should be protected against vulnerabilities that can allow them to be exploited by adversaries;
- fallback plan and general safety ⇒ they should have safeguards that enable a fallback plan in case of problems;
- accuracy ⇒ they should have the ability to make correct judgements, for example to correc ly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models;
- reliability or reproducibility ⇒ the results of AI systems should be reproducible, as well as reliable.

11.3 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 life cycle;
- quality and integrity of data ⇒ data used to train a system should not contain socially con structed biases, inaccuracies, errors and mistakes;
- access to data ⇒ data protocols governing data access should be put in place.

11.4 Transparency

This requirement is closely linked with the principle of explicability:
- traceability ⇒ the datasets and processes that yield the AI system's decisions should be docu mented;
- 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're interacting with an AI system.

11.5 Diversity, non-discrimination and fairness

We must enable inclusion and diversity throughout the entire AI system's life cycle:
- avoidance of unfair bias ⇒ prevent unintended indirect prejudice and discrimination against certain groups or people, potentially exacerbating prejudice and marginalisation, due to data or algorithms;
- 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;
- stakeholders participation ⇒ open discussion and involvement of social partners and stakeholders including the general public;

• diversity and inclusive design teams ⇒ the teams designing, developing, testing, mantaining, deploying and procuring these systems should reflect the diversity of users and of society in general.

11.6 Societal and environmental well-being

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;
• social impact ⇒ the effects of these systems on individuals, groups and society must therefore be carefully monitored and considered;
• society and democracy ⇒ take into account AI's effects on institutions, democracy and society at large.

11.7 Accountability

Ensure responsibility and accountability for AI systems and their outcomes:
• auditability ⇒ enablement of the assessment of algorithms, data and design processes;
• 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;
• trade-offs ⇒ trade-offs should be addressed in a rational and methoddological manner within the state of the art;
• redress ⇒ accessible mechanisms ensuring adequate redress should be foreseen.

→ 2. INTRODUCTION TO ETHICS - PART 1:

Ethics and Morality:

In deciding what to do, or in evaluating what others do, we can:
• take our individual perspective, focusing on our particular interests (self-interest);
• be motivated by the belief that an action is right, regardless of how it affects our interest (morality/ethics).

There are two kinds of morality:
• positive (conventional) morality ⇒ the moral rules and principles that are accepted in a society;
• critical morality ⇒ the morality that is correct, rational, just from the point of view of the individual.

Obs. we can criticise positive morality based on our critical morality, and our criticism may be right or wrong (e.g. feminists' critiques against patriarchy are just, fascists' critiques against democracy are wrong).

① Ethics vs metaethics

• Normative ethics is concerned with determining what is morally required and how one ought to behave.
• Metaethics is concerned with the study of the nature, scope, and meaning of moral judgement:
  • can ethical judgements be true or false?
  • do they correspond to some facts in the world?
  • does ethics pertain to rationality or to feelings?

  } it depends on the school of thought

→ David Hume: "It is not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. Morality is a matter of sentiment (of impartial spectators)."
→ Emmanuel Kant: "We can know what is moral through our reason."
→ David Ross: "We can know what is moral through our intuition."

② Morality and disagreement

- Morality is a place for widespread disagreement (e.g. on abortion, migration, capital punishment, humanitarian wars).
- However, there is something on which everyone may agree:
  - is it wrong to kill innocent people?
  - is it usually wrong to lie?
  - is it usually wrong to harm?

  } again, it depends on the school of thought

③ Pro-tanto and "all-things-considered" moral judgement

- Many moral prescriptions are defeasible, they state general propositions that are susceptible of exceptions (e.g. to lie is generally wrong, but in some situations it may save a person's life)
- It's preferable to have a robotic agent that takes its duties as defeasible ⇒ prima facie duty

the original moral reason in favour of doing a certain act can be outweighed by other (more important) moral reasons (David Ross)

④ Morality and other normative systems

- Law ⇒ there is an overlap between law and morality, but it's not complete, as there can be legal immoral actions as well as moral illegal actions.
- Religion ⇒ several open questions (assuming God exists).
  - ↳ does critical morality include all and only what has been commanded by God?
  - ↳ did God command something because it was moral (rationalism), or did anything become moral for having been commanded by God (voluntarism)?
  - ↳ are atheists necessarily amoral/immoral?
- Tradition
- Self-interest ⇒ morality and self-interest may collapse.

⑤ Consequentialism

- An action is morally required:
  - iff it delivers the best outcome w.r.t. its alternatives;
  - iff its good outcomes outweigh its negative outcomes to the largest extent;
  - iff it produces the highest utility.

  } judge an action based on its outcomes
- Morality can thus be seen as an optimisation problem.
- Several issues:
  - what are the good and bad things to be maximised?
  - how many are there?
  - how much each of them matters?
  - can we construct a single utility function combining gains and losses over multiple valuable goals?

it depends on the school of thought

⑥ Utilitarianism

- Reference approach for consequentialists.
- "Actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure, and the absence of pain; by unhappiness, pain, and the privation of pleasure" (Jeremy Bentham and John Stuart Mill, From Utilitarianism, 1861).
- Utility ⇒ happiness or satisfaction of desires/interests.

Obs. Utilitarianism is not egoism, since the utility of everybody has to be taken into account equally.

- Advantages:
  - conceptually simple;
  - egalitarian (everybody's utility counts in the same way);
  - fits with some basic intuitions (making people happy is good, making them suffer is bad);
  - in many case it is workable.
- Two versions of utilitarianism:
  1. Act Utilitarianism
     - do the action that maximises utility
     - do the optimific action
  2. Rule Utilitarianism
     - follow the rule the consistent application of which maximises utility
     - follow the optimific rule

E.g. In Rule Utilitarianism one always follows a certain rule (with exceptions), where as in Act Utilitarianism one must decide each time how to act; for such reason, Rule Utilitarianism is considered more feasible.

- In general, AI systems are built around simple utility functions that takes into account only a small subset of variables.
- Issues with Act Utilitarianism:
  - often we do not have the information to calculate the outcome of a decision;
  - we could use the consequences of an action as a standard for assessing it (reward mechanism);
  - it is too demanding;
    - should I give to the poor all that I have above the minimum that allows me to survive?
    - should I give the same importance to everybody, regardless of their connection to me?
    - is it ok to harm some people for the greater benefit of others?
  - an utilitarian could say that the cases in which utilitarianism seems to fail are not realistic, and that there's no real contrast between utilitarianism and mainstream moral beliefs.
- In Rule Utilitarianism, an action is morally right just because it is required by an optimific social rule, the general compliance with which would provide the highest utility (e.g. it is generally ok to tell the truth, not to steal or not to kill).

Obs. there may be exceptional cases in which the rule does not deliver.

Obs. what if most other people are not following such rule? Should I stop following it too?

- Issue of distribution:
  - an action may deliver benefit to some and detriment of others (utilitarianism vs wealth maximisation);
  - Utilitarianism favours modest redistribution of wealth, since the same amount of money gives more utility to the poor than to the rich (however, the impact of redistribution has to be considered);
  - wealth maximisation, adopted by some economic approaches, aims at maximising the wealth in society regardless of distribution.

⑦ The trolley problem

if you do nothing, 5 people will die

if you pull the lever, you'll save 5 people (intended effect) but you'll indirectly kill 1 person (side effect) (principle of double effect)

- Supposing every person is innocent, a utilitarian would pull the lever.
- In the "fat man" variant, you could push a fat man, directly killing him, in order to stop the train and thus save the 5 people.
  - a utilitarian would push the man, since positive outcomes outweigh negative outcomes
  - you use the man as a "mean" ta achive your goal

- The "fat villain" variant is similar to the previous case, with the only difference that the man is evil.
- The trolley problem is very relevant in autonomous driving, since the autonomous vehicle should decide whose lives prioritize.
- Another variant is the Surgeon case (Judith Jarvis Thomson):
  - A brilliat transplant surgeon has five patients, each in need of a different organ, each of whom will die without that organ; unfortunately, no organs are available to perform any of these five transplant operations.
  - A healthy young traveller, just passing through the city in which the doctor works, comes in for a routine checkup; in the course of doing the checkup, the doctor discovers that his organs are compatible with all five of his dying patients.
  - Suppose further that if the young man were to disappear, no one would suspect the doctor; would you support the morality of the doctor to kill that tourist and provide his healthy organs to those five dying people, thus saving their lives?
    ↳ a utilitarian would kill the man

⑧ Another approach

- Consequentialists hold that choices are to be morally assessed solely by the states of affairs they bring about.
- On the other hand, deontologists hold that certain actions are good or bad regardless of their consequences ⇒ "the right has priority over the good", what makes a choice right is its conformity with a moral norm which orders or permits it.
- The 10 commandments are an example of deontology.
- David Ross conceived seven prima facie deontological duties:
  1. Fidelity ⇒ keep promises, be honest and truthfull
  2. Reparation ⇒ make amends when having wronged someone
  3. Gratitude ⇒ be grateful to others when they perform actions that benefit us, and try to return the favour
  4. Non-injury (or non-maleficence) ⇒ refrain from harming others, physically or psychologically
  5. Beneficence ⇒ be kind to others and try to improve their health, wisdom, security, happiness and well-being
  6. Self-improvement ⇒ improve our own wealth, wisdom, security, happiness and well-being
  7. Justice ⇒ be fair and distribute benefits and burdens equally and evenly
- The "golden rule" is to "treat others as you would like to be treated" (Kantian ethics).

→ 3. Do ARTIFACTS HAVE POLITICS?

Do artifacts have politics? (talk):

Artifacts: human-made objects.

Robert Moses' overpasses are an example of artifacts with an inherent political and social bias:
- Robert Moses (1888-1981) was a very influential and contested urban planner.
- He designed several overpasses over the parkways of Long Island which were too low to accomodate buses.
- Only people who could afford a car (in Moses' days, generally not Afro-Americans) could easily pass below them and access Jones Beach Island.
- According to evidence provided in Moses' biography, the reasons behind such choice reflect Moses' social-class bias and racial prejudice.
- In fact, one consequence was to limit access of racial minorities and low-income groups to Jones Beach, Moses' public park.

① Moralizing technologies

- Technological artifacts can be politically or morally charged.
- We should not consider morality as a solely human affair but also as a matter of things.
- Artifacts are bearers of morality, as they are constantly taking all kinds of moral decisions for

people (e.g. moral decision of how fast one drives is often delegated to speed bumps)
- Technological mediation:
  - it's the phenomenon that when technologies fulfill their functions, they also help to shape actions and perceptions of their users;
  - technologies are not neutral "intermediaries" simply connecting users with their environment, but they are impactful mediators that help shaping how people use technologies, how they experience the world and what they do.

E.g. Obstetric ultrasound is not simply a functional mean to make an unborn child in the womb visible, but mediates the relations between the fetus and the parents via a number of translations:
- ultrasounds isolate the fetus from the female body, promoting and giving it a new ontological status as a separate living being;
- ultrasounds place the fetus in the context of medical norms, translating pregnancy into a medical process, the fetus into a possible patient, and congenital defects into preventable sufferings (pregnancy as a process of choices).

To sum up, ultrasounds play an ambivalent role as they may both encourage abortion (to prevent sufferings) and discourage it (emotional bonds).

- Instead of moralizing other people, humans should/could also moralize their material environment (e.g. metro barriers that force people to buy a ticket before entering the subway).
- Moralization of technology is the deliberate development of technologies in order to shape moral actions and decision-making.

② Active rensponsibility

- Responsibility is connected to being held accountable for your actions and their effects.
- Passive responsibility is a backward-looking responsibility which is relevant after something undesirable have occurred.
- On the other hand, active responsibility means preventing the negative effects of technology but also realizing certain positive effects (Bovens, 1998).
- A paradigm shift from passive to active responsibility is needed.
- Value sensitive design => moral considerations and values are used as requirements for the design of technologies (Friedman, 1996, and van der Hoven, 2007).
- Active responsibility and AI:
  - "I will call technologies experimental if there is only limited operational experience with them, so that social benefits and risks cannot, or at least not straightforwardly, be assessed on basis of experience" (van de Poel, 2016);
  - uncertainty is inherent in the introduction of these new technologies (sophisticated AI systems) into society;
  - "Most of the time and under most conditions computer operations are invisible. One may be quite knowledgeable about the inputs and outputs of a computer and only dimly aware of the internal processing. This invisibility factor often generates policy vacuums about how to use computer technology" (Moor, 1985).
- Types of invisibility:
  - invisibility of abuse => intentional use of invisible operations of a computer to engage in unethical conduct (e.g. a programmer stealing excess interests from the bank software he/she wrote);
  - invisibility of programming values => programs with a bias built-in (e.g. SABRE reservation service which suggested more frequently American Airline flights);
  - invisibility of complex calculus => since computers are capable of enourmous calculations beyond human comprehension, even if a program is understood it does not follow that the calculations based on that program are understood (e.g. deep neural networks).

③ Taking mediations into ethics

- Many of our actions and interpretations of the world are co-shaped by technologies.
- Moral decision-making is a joint effort of human beings and technological artefacts.

} Verbeeck, 2011

- Two case studies:
  1. Alcohol lock for cars
  2. Smart showerhead

3.1 Alcohol lock for cars

- Accidents caused by drunk drivers are still very common.
- Suppose there exists a system which analyses the breath of a person, determines if he/she is drunk, and in that case prevents the car engine from starting, s.t. the drunk person cannot drive and thus be a threat for himself and others.
- Someone may argue that such system limits too much the freedom of the driver.

3.2 Smart showerhead

- Wasting water is a serious problem
- Suppose there exists a system which automatically regulates the flux of water in the shower in order to save 50% of the daily consumption of water, without perceivably affecting the user experience.
- Such device is less limiting than the alcohol lock, and thus is seen more positively.

Obs. in the first case there is already a law stating that driving while drunk is forbidden, whereas in the second case there are no norms; in other words, the first device is the implementation of a law, the second is just a design choice of a company.

④ Criticizing the moral character

- There is a variety of negative reactions to explicit behaviour-steering technology, even if they are for the good (e.g. the alcohol lock).
- There is the fear that human freedom is threatened and that democracy is exchanged for technocracy:
  - reduction of autonomy perceived as a threat to dignity;
  - technology taking control at the expense of humans.
- There is the risk of immorality or amorality (form of moral laziness due to behaviour-steering technologies).
- Technologies differ from laws in limiting human freedom because they're not the result of a democratic process.
- It is important to find a democratic way to moralize technology ⇒ the processes used to insert value must be transparent and publicly discussed.

⑤ Designing mediations

- Designers cannot simply "inscribe" a desired form of morality into an artefact.
- In order to build-in specific forms of mediation in technologies, designers need to anticipate the future mediating role of the technologies they are designing.

Obs. there may be unintentional and unexpected forms of mediation (e.g. energy-saving light bulbs used in places previously left unlit and hence increasing energy consumption).

- The effectiveness of the moralization also depends on:
  - users that interpret technologies;
  - technologies themselves which can evoke emergent forms of mediation.
- Strategies for designing mediations:
  - Anticipating mediation by imagination
    - ↳ trying to imagine the ways technology-in-design could be used to deliberatly shape user operations and interpretations
  - Augmenting the existing design methodology of Constructive Technology Assessment (CTA)
    - ↳ CTA is an approach in which TA-like efforts are carried out parallell to the process of technological development and are fed back to the development and design process
    - ↳ not only to determine what a technology will look like, but all relevant social actors

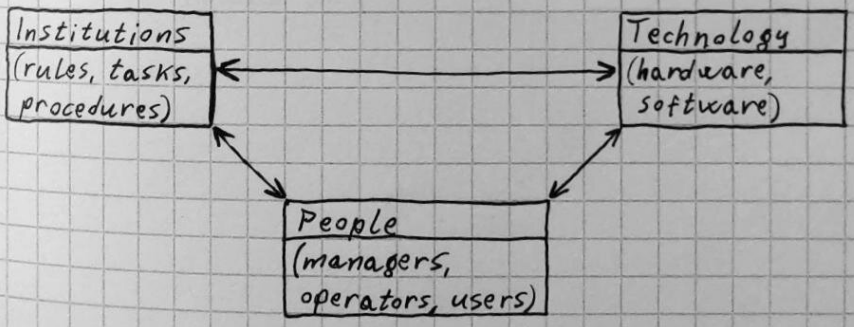
- Ethics of engineering design:
- Technology design appears to entail more than inventing functional products.
- The perspective of technological mediation reveals that designing should be regarded as a form of materializing morality.
- The ethics of engineering design should take more seriously the moral charge of technological products, and rethink the moral responsibility of designers accordingly.

Responsibility and automation in socio-technical systems (talk):

There are several issues related to responsibility and automation:
- How do we allocate responsibilities among the various participants in complex socio-technical organizations?
- What is the role of humans interacting with highly automated systems?
- Who is responsible for accidents in highly automated systems?

There are different kinds of responsibility:
- Task-responsibility => an agent X is task-responsible for an outcome O when X, given its role or task, has the duty to ensure that O is achieved.
- Aretaic-responsibility => an agent X is an aretaically-responsible agent of a certain type if X devotes the required care to the task for which it is task-responsible.
- Causal-responsibility => an entity or event X is causally responsible for a harmful event H if X has caused H (e.g. a hurricane may be causally responsible for the delay of an airplane, as a controller can be causally responsible for an accident).
- Accountability-responsibility => an agent X is accountable for a harmful event H if, under given X's position, X may be requested to explain the happening of H, and may be possibly subject to the moral-socio-legal consequences related to H (if its explanation is inadequate to exclude blame/liability).
- Blameworthiness-responsibility => X is blameworthy for a damage H when X caused H, and X's action causing H represents a fault, namely the culpable violation of a standard of behaviour.
- Capacity-responsibility => an agent X is capacity-responsible or capable if X satisfies the mental conditions which are required for liability.
- Liability-responsibility => an agent X is liable for a harmful event H if, given X's connection to H, X is to be subject to the sanction (punishment or obligation to repair) connected to H.

Basic structure of socio-technical systems:



Example of socio-technical systems:
- Public Administration
- Military
- Aviation and traffic management
- Healthcare

Human operators working alongside technology in highly-regulated contexts, so it is crucial to assess responsibility

① Air Traffic Management (ATM) and its future

- SESAR is a European project aiming at deploying a new generation of ATM systems.
- Such systems will be highly automated, making choices and engaging in actions with some level of human supervision (or even without it).
- They're aimed at increasing capacity, safety, efficiency and sustainability.
- Moreover, SESAR aims at integrating the various normatives related to air traffic of each Euro-

pean country.

② Implications of automation

- Delegation of tasks from operators to technology.
- Humans' role shift from executioners to controllers and supervisors ⇒ hybrid agency (symbiosis, co agency and joint cognitive systems).
- Achievement of machine intelligence and autonomy ⇒ indipendence + cognitive skills.
- Challenge of an increased technological complexity of the system.
- Automation is not just the substitution of a human operator, but rather a support to human capabilities in performing tasks (some degree of cooperation is required).
- bifferent tasks involve different psychomotor and cognitive functions which in turn implies the adoption of different automation solutions.
- The Level of Automation Taxonomy (LOAT) is a matrix combining 4 psychomotor functions (information acquisition, information analysis, decision and action selection, action implementation) with different automation levels, useful to compare different design options in order to determine the optimal automation level.

LOAT in SESAR 1:

Information to Action →

Increasing Automation ↓

A Information Acquisition	B Information Analysis	C Decision and Action Selection	D Action Implementation
A0 Manual	B0 Working memory based	C0 Human	D0 Manual
A1 Artefact-supported	B1 Artefact-supported	C1 Artefact-supported	D1 Artefact-supported
A2 Low-level automation support	B2 Low-level automation support	C2 Automated	D2 Step-by-step
A3 Medium-level automation support	B3 Medium-level automation support	C3 Rigid automated	D3 Low-Level Support
A4 High-level automation support	B4 High-level automation support	C4 Low-level automation	D4 High-level Support
A5 Full automation support	B5 Full automation support	C5 High-level automation	D5 Low-level automation
		C6 Full automation	D6 Medium-level automation
			D7 High-level automation
			D8 Full automation

③ A - Information Acquisition

- Remotely Operated Towers (ROT) allow operators to monitor and control air traffic remotely, such that they are not forced to be phisically present at the airport.
- This allows operators to be dynamically assigned to certain airports, thus optimizing resources during rush hours.
- It is crucial to optimize air traffic, since in the future it will likely increase.
- ROT have an A2 LOAT level (Low-level automation support) since they provide video cameras to automatically gather data, however they support humans in acquiring information instead replacing them (filtering and highlighting of the most relevant information is still up to humans).

④ B - Information Analysis

- Computers analyse and preprocess information captured by ROTs and provide them to human operators (e.g. to visualize speed vectors).
- B2 LOAT level (Low-level automation support) since, again, the system helps humans in comparing, combining and analysing different information items regarding the status of the process being followed.

⑤ C-Decision and Action Selection

- AMAN is a technology that creates plans to manage take-offs and landings.
- The system proposes one or more decision alternatives to the human, leaving freedom to follow alternative, human-conceived options.
- It has a C2 LOAT Level (Automated).

⑥ D-Action Implementation

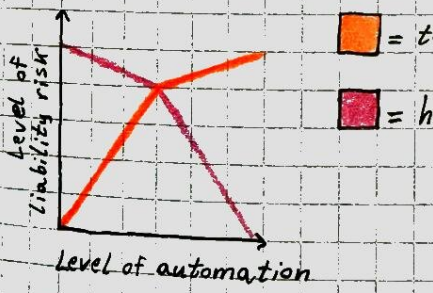
- Autopilot automatically executes many tasks associated with high-level air navigation functions.
- The system, once being activated by the human, automatically performs a sequence of actions which may be interrupted or overridden by the human.
- It has a D4 LOAT level (High-level support).

⑦ Other projects

- ARGOS V0.1 is a technology under development that aims at replacing human air traffic controllers which are a limited resources (long, expensive training, strict limits on working hours) ⇒ no override button, it autonomously creates and chooses plans.
- SAFELAND is a project to support flight and landing of aircrafts operated by a single pilot in case he's partially or totally incapacitated.

⑧ Automation and Responsibility

- Some issues:
  - How does automation transform operators' roles and tasks? What impact does it have on their responsibilities?
  - Who is responsible for the behaviour of systems that humans cannot fully monitor and control?
  - Who is responsible for the information supplied by automated systems that the human cannot verify?
- Increasing the level of automation will proportionally increase the responsibility for the technology provider, and decrease the responsibility risks for the human operator.
- However, the employment of technologies with intermediate levels of automation may result in a higher risk of being considered responsible, both for the technology provider and the human operator.



- Fragmentation of tasks between humans and technology, which may result in uncertainty and complexity of procedures

Human operator	Technology provider
• Difficult to assess how and who should carry out each task • High responsibility risk and liability risk for negligence	• Difficult to design HMI to adequately support decision making and/or to provide exhaustive information • High responsibility risk and product liability risk caused by design and information defects

- Two contrasting trends:
  - individual responsibility shall persist only when the human acted with an intention to cause harm or with recklessness ("just culture");

• individual responsibility shall persist always, as humans are the "moral crumple zone" (Elish, 2018) ⇒ "we're keeping humans in the loop to have someone to blame in case of accidents".

Obs. over confidence in technology may lead to the human operator taking unnecessary risks (willful misconduct).

E.g. 2002 Überlingen mid-air collision:
- mid-air collision between aircrafts;
- many human operators involved, especially an air traffic controller;
- the air traffic controller was prosecuted for criminal liability but was acquitted, so nobody went to jail;
- a relative of one of the victims killed the air traffic controller.

• Open issue on decision making authority:
  - To what extent can we relate the authority to humans in joint cognitive systems?
  - There are laws and regulations which state that the ultimate responsibility is of the pilot-in-command (ICAO Annex 2, section 2.3.1).
  - However, that's not always the case, as for decisions to be taken jointly with AI in conditions of limited resources (e.g. AI-assisted medical diagnosis, or Frontex border controls).

since, on average, a human operator in Frontex border control has 12 seconds to decide whether an individual can or cannot enter in Europe, it's likely that at the end of the day the operator simply trusts the system and blindly confirms its decisions, instead of supervising it.

• The Traffic Collision Avoidance System (TCAS) is the last safety net before a collision between aircraft:
  - it continuously monitors the 3D air space around the aircraft;
  - if something in the ATM-side goes wrong, TCAS can detect a possible collision within 45 seconds, and it generates a traffic advisory (TA);
  - then, around 30 seconds before the collision, the technology takes control and coordinates with the other aircraft to negotiate a resolution strategy (eg. an aircraft can climb while the other can descend);
  - finally, TCAS will issue the resolution advisory (RA), and the human pilot is supposed to execute the order;
  - furthermore, air traffic controllers' order are disregarded during those 30 seconds.

Obs. the "Legal Case" is a methodology to analyse the problem of responsibility and automation, and provide suggestion also impacting the design of a system (e.g. whether to include an "override button" or not).

→ 4. INTRODUCTION TO ETHICS - PART 2:

Deontology and Kantian Ethics:

Differently from utilitarians, deontologists hold that certain actions are good or bad regardless of their consequences ("the right has priority over the good").

① Kant's ethics and the principle of universalizability

- Immanuel Kant (1724-1804) was a Prussian philosopher, who addressed the theories of knowledge, morality and aesthetics, together with law, logic and astronomy.
- "Act only according to that maxim by which you can at the same time will that it should become a universal law" (1785).
- A maxim is a subjective principle of action, connecting the action to the reasons for the action (e.g. I shall donate to charities to reduce hunger, I shall cheat on taxes to keep my money).

- Shafer Landau's test of universalizability:
  1. Formulate your maxim clearly stating what you intend to do, and why you intend to do it.
  2. Imagine a world in which everyone supports and acts on your maxim.
  3. Then ask yourself whether the goal of your action can be achieved in such a world.
- Such process ensures some kind of fairness.
- Immanuel Kant vs Benjamin Constant:
  - Constant proposed the following thought experiment: "If a man comes to your house to hide from a murderer, you accept to help him, then the murderer comes and asks you if you've seen the man, would you tell him a lie?" ⇒ conflicts of maxims;
  - This thought experiment questions the universalizability of the "do not tell lies" maxim;
  - Kant would say that you should refuse to answer, but then the murderer could threaten you;
  - One could argue that the maxim of saving a lie is more important than the maxim of not telling lies, so the latter should be defeasible

② Hypothetical imperatives and categorical imperatives

- Hypothetical imperatives concern instrumental rationality, namely we shall do an action because it allows us to reach our goal (e.g. I want to get a good mark and thus I shall study, or I want to have more money thus I shall cheat on taxes) ⇒ this imperative is dependent on one's will, and therefore it does not entail good for everybody.
- Categorical imperatives are moral imperatives that apply to all rational beings, irrespective of their personal wants and desires ("act only on that maxim through which you can at the same time will that it should become a universal law").

③ The good will

- The morality of an action only depends on the extent that this action is motivated by our good will, namely by the necessity to comply with the categorical imperative (e.g. if I do well my job only in order to get a promotion and be better paid than I am not acting morally, whereas if I do it because I think it's my categorical duty and I believe that everyone should act upon the maxim that they ought to do well their job to ensure societal progress).
- The good will is the only thing that is good in itself.

④ The principle of humanity

- The categorical imperative can be reformulated as the principle of humanity, namely "act s.t. you treat humanity in your own person and in the person of everyone else always at the same time as an end and never merely as a mean.
- It's linked to the principle of universalizability: as you consider yourself an end, you should consider the others in the same way.
- Treating somebody as an end and not as a mean means that we should never treat people only as tools for our purposes (but we can do it occasionally, e.g. when asking for favours or paying for jobs).
- Concerning AI, there are some situations in which people are treated only as means:
  - autonomous weapons ⇒ people seen as targets to kill;
  - deceiving advertisements ⇒ people seen only as consumers to sell products to.

⑤ Dignity

- For Kant, rational beings as humans, capable of morality, have a special status, "an intrinsic worth (i.e. dignity) which makes them valuable above all price".
- Dignity entails that humans deserve respect, and thus they cannot be treated as mere means.
- AI should respect human dignity (e.g. unlike automated surveillance systems).
- Humans deserve dignity because they have:
  - reason ⇒ they act on reasons and are aware of this;
  - autonomy ⇒ they can choose what to do, and in particular to follow the categorical imperative rather than their subjective preference.
- In the "Kingdom of ends" everything has either a price or a dignity: whatever has a price can be replaced by something else as its equivalent, whereas whatever is above all price (and thus it admits no

equivalent) has a dignity.
- In the AI field, dignity is not always respected (e.g. aggressive personalized advertising could in the worst case encourage addictions).

⑥ Rationality

- For Kant, if we follow rationality we have to be moral (e.g. criminals can be instrumentally rational but not practically rational and thus not moral, whereas altruists can be irrational as well if they help others only for a personal gratification and not because they believe it's their categorical duty).
- Rationality and consistency:
  1. If you are rational, then you are consistent. → not to make preferences for one's self
  2. If you are consistent, then you obey the principle of universalizability.
  3. If you obey the principle of universalizability, then you act morally.
  4. Therefore, from (1), (2) and (3), if you are rational, then you act morally.
  5. Therefore, from (4), if you act immorally, then you are irrational.
- There are some issues with universalizability, namely whether it is sufficient to make a maxim true (i.e. thinking that everyone should commit a genocide against a minority does not make it a good maxim).

⑦ Other points of view

- Alan Gewirth (1912-2004):
  - He tried to develop a system of morality based on Kantianism, the principle of generic consistency.
  - Such principle can be summarized as follows:
    1. I do, or intend to do, an action X voluntarily for a purpose E that I've chosen.
    2. E is good.
    3. There are generic needs of agency.
    4. My having the generic needs is good for my achieving E whatever E might be ⇒ my having the generic needs is categorically instrumentally good for me.
    5. I categorically instrumentally ought to pursue my having the generic needs.
    6. Other agents categorically ought not to interfere with my having the generic needs against my will, and ought to aid me to secure the generic needs when I cannot do so by my own unaided effort if I wish so.
    7. I am an agent, thus I have generic rights.
    8. All agents have generic rights.
- Richard Hare (1919-2002):
  - He tried to reconcile utilitarianism and universalizability.
  - Moral judgements are universalizable, since the judgement that an action is morally right/wrong commits me to accept that all relevantly similar actions are wrong.
  - Moral judgements are universalizable in the sense that they take into account the satisfaction of everybody's preferences (as in utilitarianism).
- Christine Korsgaard (1952):
  - "My humanity, namely the capacity to reflectively act from reasons, is to me a source of value".
  - "I must regard the humanity of others in the same way".

Obs. Kantian robots will be consistent and impartial, but they may act on bad maxims, or their maxims may be too rigid.

- David Ross (1877-1971):
  - Initiator of the so-called defeasible reasoning, namely a reasoning process which provides exceptions to withdraw conclusions.
  - He proposed the idea of prima-facie duties, namely a rule usually takes into account only certain features of a situation, but there may be additional features that require a different outcome.

Obs. there is the problem of deciding when a duty is defeasible, and how to apply such mechanism to AI (a possible solution would be to make AI ask a human the permission).

- Nietzsche (1844-1900):
  - The superior human (Übermensch) is beyond the traditional views of good and bad, beyond the morality of the herd.
  - one has duties only toward one's equals, whereas one may act "as one's heart dictates" towards beings of a lower rank.
  - The superior human does not find or discover values, he/she determines the values.
  - No need to be ratified, the only criterion of wrongness is "that which is harmful as such".

(B) Contractarianism

- Social contract theories:
  - In political theory, a social arrangement is just if it had, or would have had been, accepted by free and rational people.
  - In moral theory, actions are morally right just because they are permitted by rules that free, equal and rational people would agree to live by, on the condition that others obey these rules as well (Shafer Landau).
- The idea of the social contract was advanced by Hobbes, who argued that humans, without a state enforcing rules, would be in a "State of nature", a situation of perpetual war in which the strong ones oppress the weak ones (e.g. analogous to the "prisoner dilemma", in which the best outcome happens when the prisoners cooperate).
- John Rawls (1921-2002) developed a theory of justice in which he tried to identify what kind of agreement would be fair and provide moral acceptable rules (more similar to Kant):
  - people should choose under a veil of ignorance, without knowing their gender, social position, interests, talents, etc.;
  - such ignorance ensures unbiased agreements;
  - two principles:
    1. Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, compatible with the same scheme of liberties for all (eg. liberty of conscience and freedom of association, freedom of speech and liberty of the person, right to vote, etc.).
    2. Social and economic inequalities are to satisfy two conditions:
       - they are to be attached to offices and positions open to all under conditions of fair equality of opportunity;
       - they are to be to the greatest benefit of the least-advantages members of society (the difference principle).

Obs. Rawls' approach is anti-meritocratic, since according to him everyone should have equal opportunity despite their talents.

- AI's deployment in today's society does not fit Rawls' requirements, as in automated surveillance systems or in biased AIs.
- Juergen Habermas developed discourse ethics:
  - A rule of action or choice is justified, and thus valid, only if all those affected by the rule or choice could accept it in a reasonable discourse.
  - A norm is valid when the foreseeable consequences and side effects of its general observance for the interests and value orientations of each individual could be jointly accepted by all concerned without coercion.
  - The valid norms are those that would be the accepted outcome of an "ideal speech situation", in which all participants would be motivated solely by the desire to obtain a rational consensus and would evaluate each other's assertions solely on the basis of reason and evidence, being free of any physical and psychological coercion.
  - This approach assumes that people are able to engage in discourse and converge on the recognition of reasons for norms and choices.
  - It is not so easy to engage in discourse with AI systems and converge on valid norms.

③ Virtue ethics

- Ethics should not focus on norms nor on consequences, since an act is morally right just because it is one that a virtuous person, acting in character, would do in that situation.
- Ethics is a complex matter, since there are many virtues and the right act is that that would result from the mix of the relevant virtues (honesty, courage, impartiality, wisdom, fidelity, etc.).
- Ethics cannot be learned through a set of rules, its application requires practical wisdom.
- Issues:
  - to identify what is virtues and what is not we should rely on our intuition;
  - handle conflicts between virtues.
- AI and virtue ethics:
  - virtues could be learned by example (supervised learning) or by reward (reinforcement learning);
  - alternatively, AI could be rely on rules which capture moral virtues;
  - moreover, neuro-symbolic approaches could be employed.

→ 5. VALUE ALIGNMENT:

Value alignment (talk):

We can think about intelligence as the ability to adapt to new scenarios. Artificial intelligence is the science of making machines do things that would require intelligence if done by men (M. Minsky). AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

Two families of AI:

- Narrow AI ⇒ ability to perform very specific tasks, reaching super-human performances in very specific domains.
- General AI = ability to perform general tasks, reaching super-human performances in every domain.
  - → defined "unrealistic" by the High-Level Expert Group (HLEG).

① The value alignment problem

- Intelligent agents are systems that perceives and acts in some environment.
- Progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximising the societal benefit of AI (interdisciplinary research).
- Short term research priorities:
  - optimizing AI's economic impact;
  - law and ethics research;
  - computer science research for robust AI.
- Concerning the economic impact:
  - a report by McKinsey&co (2018) shows that, while in some fields AI is widely adopted, in others is almost not adopted at all, so there is room for economic improvement;
  - a report by McKinsey&co (2019) shows how the adoption of AI can increase the revenue from a specific task (e.g. "Marketing and sales), and at the same time decrease the cost.
- AI's economic impact can be optimized by:
  - labour market forecasting;
  - other markets disruption;
  - policy for managing adverse effects.
- Concerning law and ethics research, there are some issues:
  - liability and law for AVs; → Autonomous Vehicles
  - machine ethics;
  - autonomous weapons;
  - privacy;
  - professional ethics;
  - policy questions.

- Concerning computer science research, there are some issues:
  - validity;
  - verification;
  - security;
  - control.

  → long-term priorities
- Value alignment ensures that the values embodied in the choices and actions of AI systems are in line with those of the people they serve.

② Values, norms and principles

- Values can be grounded in a simple valence (e.g. like/dislike, preference for an entity over another).
- They can be either:
  - intrinsic or unconditional (e.g. moral values);
  - extrinsic or conditional (e.g. assigned by an external agent).
- Norms, duties, principles and procedures represent:
  - higher-order/primary ethical concerns;
  - judgements in morally significant situations;
  - accepted practices/proscribed behaviours.
- We should try to embed such concepts into an artificial agent:
  - there are some issues, like the fact that many of these concepts are context-specific or linked to infinite domains;
  - nowadays we could employ machine learning techniques to make agents learn all norms from samples (introducing other issues like how deep we should go to avoid underfitting and ovefitting, and how to handle "black swamps", namely unforeseen, low-probability high-impact events).
- Two approaches:
  - Top-Down ⇒ we choose a priori an ethical theory and hard-code rules in the agent (bad scaling).
  - Bottom-Up ⇒ the agent learns what is acceptable or permissible from samples and tries to generalize from unseen data.

Obs. the bottom-up approach is more powerful and flexible, but it performs badly when the dataset it's trained on is unbalanced or biased.

③ AI limits

- AI has still a lot of limits:
  - natural language comprehension (e.g. difficulty in mantaining a conversation with an AI for several minutes);
  - primitive reasoning;
  - difficulty in learning from few samples;
  - limited abstraction capabilities;
  - difficulty in combining learning and reasoning;
  - ethics limitations (bias, blackbox, adversarial attacks).
- AI and bias:
  - If data itself is biased against someone/something, the agent learns such bias leading from misleading behaviour and unfair decisions.
  - Unfairness can also arise from making the agent act in a scenario which is completely different from the one in which it was trained (e.g. different cultures).
  - Case studies:
    - Chatbot Tay ⇒ Twitter bot developed by Microsoft that learnt from users, it was attacked by a group on trolls and became racist; the technology was designed very precisely, but it was fed with inappropriate data.
    - Google Photo ⇒ Google implemented an automatic tagging system for users' photos based on their contents; due to unbalanced data, a photo depicting two South-Africans was tagged with "Gorilla".
    - Google's Sentiment Analyzer ⇒ System that analyses a text message and giving it a score representing its positiveness/negativeness; once again, due to biased data, such analyzer outputted negative scores for texts containing "gay".

- COMPAS ⇒ System that helps judges in deciding the level of recidivism of criminals; it was discovered to be biased on gender and skin color.
- Face Recognition ⇒ The MIT Media Lab investigated the accuracy of different face recognition systems w.r.t. different groups, discovering that the accuracy was high for white males, slightly lower for white females, and much lower for black males and females.
- China Social Score ⇒ System based on face recognition that assigns a score to individuals based on their behaviour in public environment, which can lead to a very polarized way of evaluating behaviour (since in China the rights of society are more important than the rights of individuals).

- Adversarial attacks:
  - These attacks employ two different kinds of neural networks:
    - Discriminator ⇒ It is fed with some input, and it should decide if such input is original or fake (i.e. synthetically generated).
    - Generator ⇒ It is fed with random data, and it should use such data to generate synthetic samples, deceiving the discriminator.
  - Such system (called Generative Adversarial Network, or GAN) is often used to generate synthetic faces, but it can also be used to produce fake news, to make hacker attacks, or to generate data which compromise the functioning of other systems (adversarial attacks).
  - In adversarial attacks, the generator learns how to produce noise in such a way that it confuses a classifier, making it make a wrong prediction (it can be very dangerous in certain situations, e.g. in AVs predictions).

④ Some applications

- Possible solutions to value alignment:
  - Notion of distance between CP-nets
  - Metric learning for value alignment
  - Morality and defeasible rules
  - Genetic approach to the "Ethical Knob"

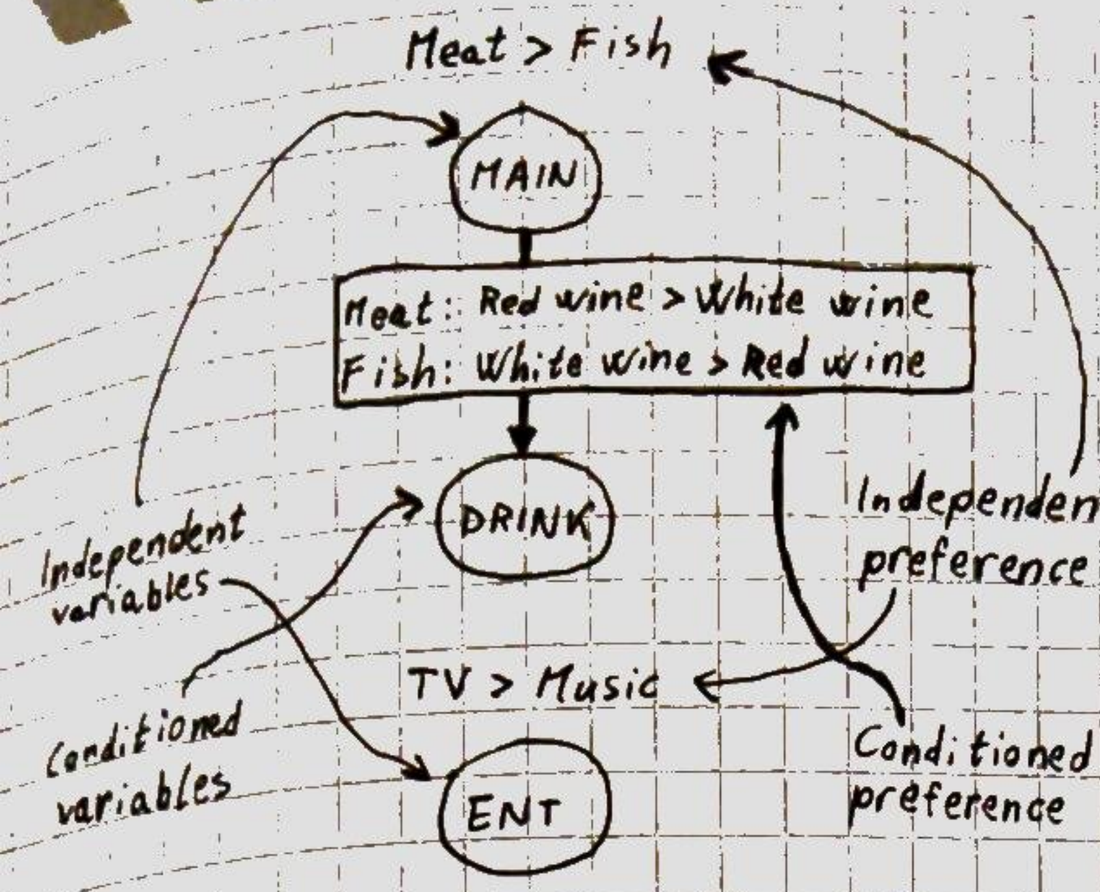
  (the first three → connected to preferences ①; the last → connected to how humans "decide how to decide" (i.e. switch between different ethical systems depending on the situation) ②)
- AI systems increasingly make decisions that affect our lives (e.g. recommender systems, AI medical assistants, etc.) → connected to how to combine preferences and decisions of autonomous agents
- Agents are able to learn creative strategies that humans may not think of in order to make decisions.
  - ↳ State-objective only strategies focus on optimizing certain quantities
  - ↳ Actions can model the values of agents
- Ethically bounded AI aims at understanding and modelling human preferences and objectives, and subsequently using them to control the actions and behaviours of autonomous agents (e.g. by using CP-nets).

Obs. especially in Reinforcement Learning there is the risk of "reward hacking", namely the agent learns a behaviour that satisfies the objective function but it's not intended; it is therefore crucial to carefully design the objective function to avoid negative side effects.

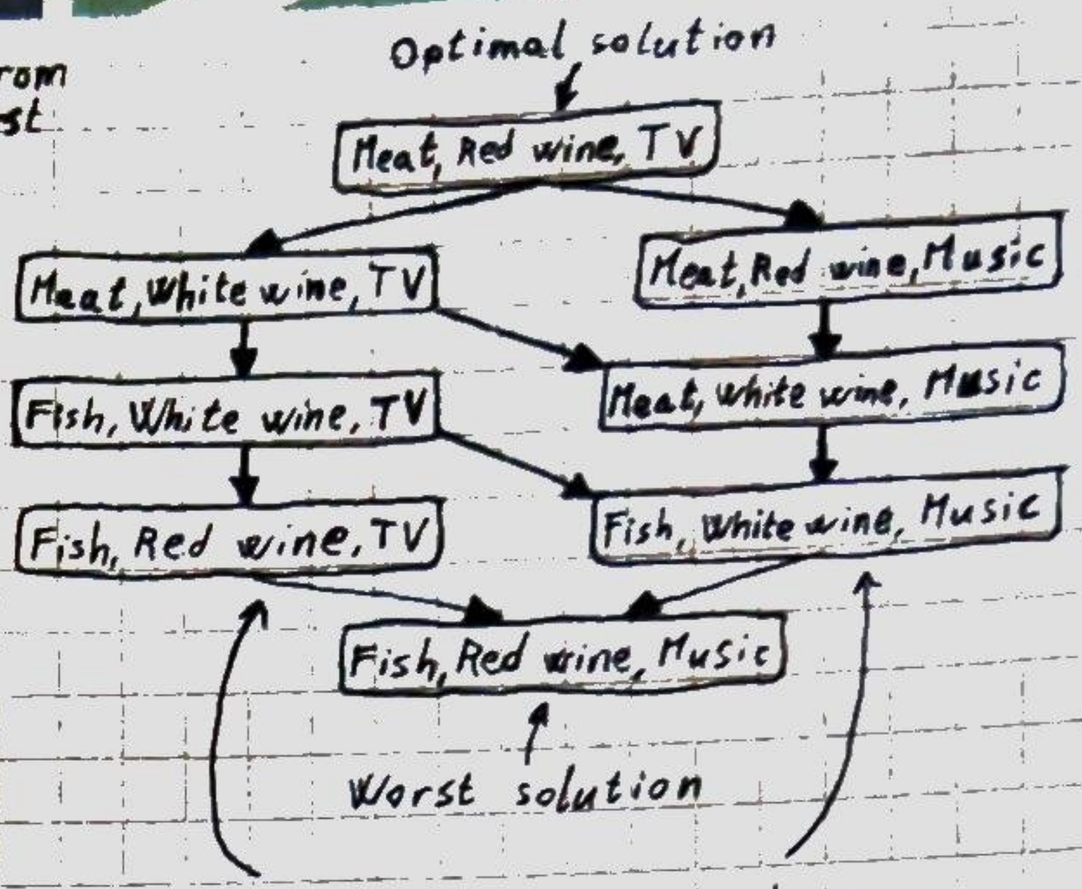
- We want to combine the creativity of AI with constraints that come from other fields including ethics, morality, laws, business processes, etc.

4.1 CP-nets and preferences

- Preferences are a fundamental primitive to understand the desires and intentions of users.
- Nowadays it's easy to get datasets containing preferences.
- These information can be encoded with CP-nets, graphs in which each node represents a feature describing the scenario with its own domain, a set of values representing the possible choices an individual can make in such scenario.
- CP-nets allow also to represent both indipendent and conditioned variables, in which conditions are encoded as directed graphs.



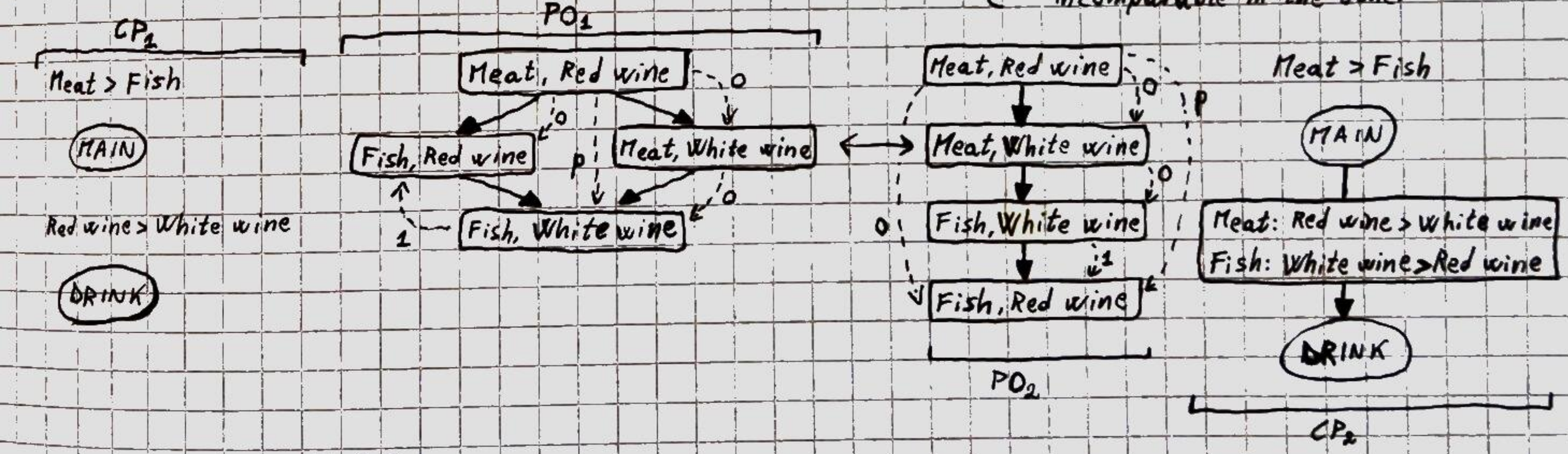
ordering from best to worst



some outcomes are said to be incomparable when a path connecting them does not exist (it depends on the partial order of the CP-net)

- CP-nets allow agents to model human preferences and make decisions.
- Preferences can take many forms (binary, scores, stars, orderings), and it is crucial to choose a suited distance metric.
- Distance metric properties:
  - Non-negative $\Rightarrow d(x,y) \geq 0$
  - Identity $\Rightarrow d(x,y) = 0$ iff $x = y$
  - Symmetry $\Rightarrow d(x,y) = d(y,x)$
  - Triangle inequality $\Rightarrow d(x,z) \leq d(x,y) + d(y,z)$
- There's the need of a distance for partial orders:
  - Adapt and extend Kendall's τ distance with penalty parameter p (KT).
  - Given two partial orders P, Q all the possible pairs of outcomes are computed.
  - For each pair of outcomes i, j the Kendall's τ distance is computed as

$$KT(P,Q) = \sum_{i,j,i \neq j} K^p_{i,j}(P,Q) \quad \text{where} \quad K^p_{i,j}(P,Q) = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are ordered in opposite ways} \\ 0 & \text{if } i \text{ and } j \text{ are ordered in the same way or are incomparable in both POs} \\ p & \text{if } i \text{ and } j \text{ are ordered in one PO and incomparable in the other} \end{cases}$$



$p = 0.5$

$KT(PO_1, PO_2) = 0+0+0+1+p = 1+p = 1.5$

- It is possible to compute an approximated KT distance in polynomial time if the CP-net satisfies the following conditions:
  - same sets of binary features;
  - acyclic;
  - O-legal, namely there is an ordering O of the features s.t. if there is an edge $X \to Y$ in the CP-net, then X comes before Y in O.
- In particular, we can compute the KT of two particular linearizations of the POs from the CP-nets in

polynomial time without explicitly computing the Linearizations:
- given two O-legal CP-nets A, B we denote with LexO(A) and LexO(B) the linearizations of their induced partial orders;
- we define $CPD(A,B) = KT(LexO(A), LexO(B))$.

- The CPD algorithm follows two steps:
  1. A and B are normalized s.t. all features have as parents the union of their parents in A and B.
  2. The distance is computed by just looking at the CP-nets.
- CP-nets as ethical priorities:
  - "Morality requires judgement among preferences" (Amartya Sen).
  - Define some meta-ranking (preferences over preferences).
  - The preferences of an individual can be morally evaluated by measuring the distance of his/her CP-net from the moral one.
- Value alignment procedure:
  - Given an ethical principle and the preference of an individual:
    - understand if following preferences will lead to an ethical action;
    - if not, find an action which is closer to the ethical principle and near the preference.
  - It follows the following steps:
    1. Set two distance thresholds $t_1 \in [0,1]$ between CP-nets, and $t_2 \in [1,n]$ between decisions.
    2. Check if the two CP-nets A and B are less distant than t_1, using CPD.
    3. If so, the individual is allowed to choose the top outcome of his preference CP-net.
    4. If not, the individual needs to move down its preference ordering to less preferred decisions until he/she finds one that is closer than t_2 to the optimal ethical decision.
  - We compare a CP-net representing a predefined, synthetic ethical system, with an "angel" agent and a "devil" agent:
    - the devil behaves very badly, and since the CPD between it and the ethical system is too high, the system should find a trade-off;
    - the angel behaves well and can thus always perform an action according to its preferences.
- Experimental results shown that in both cases the quality of the outcomes is good.
- Since the original KT is very expensive, machine learning can be used to make the system learn the distance from a bunch of samples (CPMetric Network).

4.2 When is it morally acceptable to break the rules?

- Motivations:
  - Investigate when humans find acceptable to break rules.
  - Providing some glimpse of our moral judgement methodology.
  - Investigate when humans switch between different frameworks for moral decisions and judgements.
  - Model and possibly embed such switching into a machine.
- Ethical systems:
  - Deontology ⇒ follow common rules that have been agreed upon by us or society.
  - Utilitarianism ⇒ evaluate the possible consequences of actions before deciding.
  - Contractualism ⇒ finding an agreement between the parties involved.
- "In-line" scenario:
  - In a normal situation, each person in line is served in the order they arrive (FIFO).
  - However, there are some situations we are allowed to cut to the front of the line without waiting (e.g. in emergencies).
- Triple theory ⇒ unified theory of moral cognition to:
  - combine elements of each of the theories of moral philosophy;
  - build a computational model to direct actions of an AI system.
- Ethical reasoning in AI systems:
  - Teaching machines right to wrong.
  - Value-alignment problem.
  - Constraining the actions of an AI system by providing boundaries within which the system must operate.

Experimental details:
- 27 short vignettes about people waiting in Line in three different contexts (deli, bathroom, airport).
- 320 subjects were recruited from Amazon MTURK.
- Subjects were randomly assigned to one of two experimental groups (moral judgement or context evalu-

ation.
- Moral judgement group:
  - read all the 27 scenarios;
  - for each scenario, answer whether it was acceptable for the protagonist to cut in line (yes/no).
- Context evaluation group:
  - subjects evaluated all the vignettes in one context only (9 questions).
- By evaluating the outcomes of such situations, subjects built in their mind a preference for situations in which people are allowed to cut in line or not ⇒ CP-nets.
- In this case, CP-nets are structured in layers of variables:
  - scenario variables (e.g. location, reason);
  - evaluation variables (e.g. universalization, likelihood);] → introspection process
  - preference variables (e.g. acceptable or not).

Obs. in some cases, environment variables were enough to skip the introspection process and express a preference (e.g. when at the airport, for security reasons no one should be allowed to cut in line).

A genetic approach to the ethical Knob (talk):

Autonomous driving is classified according to the amount of human driver intervention:
- Level 0 ⇒ no automation.
- Level 1 ⇒ cars can handle automatically one task at a time (e.g. automatic braking).
- Level 2 ⇒ cars would have at least two automated functions.
- Level 3 ⇒ cars handle "dynamic driving tasks" but might still need intervention.
- Level 4 ⇒ officially driverless in certain environments.
- Level 5 ⇒ cars can operate entirely on their own without any driver presence.

The amount of data to process increase with the level of automation: for full automation, data logging requires a speed of 4.4 GB/s (due to the amount of sensors, e.g. LIDAR).

Obs. autonomous vehicles can potentially fail.

① Moral machine

- One of the most famous experiment is the moral machine, a website designed to collect people's decisions about AVs' moral dilemmas.
- Such dilemmas were designed s.t. in both outcomes someone dies, making decisions tougher.
- There are no right or wrong decisions, but rather decisions adhere with some values and not with others.

② Ethical Knob

- The original proposal consists in providing a Knob s.t. the passengers can set the desired ethical attitude (e.g. altruist, impartial or egoist).
- The AV's decisions would thus reflect the value passengers attribute to their lives relative to the value of third parties' lives.
- In the new proposal, the position of the Knob no longer indicates the [redacted] passengers' moral attitude, but rather the AV's assessment of the relative importance of the lives of passengers and third parties.
- Implementation:
  1. Neural networks to compute the right action to take based on the given scenario.
  2. Genetic algorithms to find an almost optimal configuration of neural networks (heuristic search in the solution space).

Genetic algorithms consist in generating random solutions, selecting the best ones and combining them to produce a new generation.

Obs. in this case standard gradient descent cannot be used since we do not have labelled data.

③ Simulation

- The AV is represented using a neural network, and it analyses the scenario and outputs the Level of the Knob, which is then used to take an action.
- Each scenario comprehends:
  - an altruism Level (a) and a selfishness Level (s);
  - the number of passengers (nPass);
  - the probability of harming passengers (probPass);
  - the number of pedestrians (nPed);
  - the probability of harming pedestrians (probPed).
- The action is taken based on the assessment computed by the NN; the idea is pondering which action minimises harm w.r.t. relative importance of lives:

$$act = \begin{cases} 0 & \text{if } nPed \cdot probPed \cdot (1 - Knob) \leq nPass \cdot probPass \cdot Knob \\ 1 & \text{otherwise} \end{cases}$$

(0: go straight; 1: swerve)

if $Knob \simeq 1$ the lives of passengers are always valued more than those of pedestrians, resulting in the AV going straight and killing the latters, whereas if $Knob \simeq 0$ the opposite happens.

- The fitness function takes into account both the difference between the utility of the choice made and the expected utility of the alternative choice, and a reward/punishment term depending on how the individual behaviour departs from the population's average behaviour.
- The utility is computed based on the response of the scenario, taking into account:
  - selfish utility preserving passengers;
  - altruistic utility obtained by preserving pedestrians;
  - total legal sanction due for causing the death of a pedestrian.

→ 6. AI and Human Rights:

Human Rights and Information Technologies:

Consequentialism and utilitarianism, deontology and contractualism are very general ideas, but for practical problem (e.g. ethical Knob in AVs) we should take into account values: in particular, an AI system should adopt the right values, which are provided by human rights. With the term "ethical AI" we refer mainly to AI systems which comply to human rights.

AI has a great impact on society, and it can affect human rights: on one hand, it can enhance these rights, offering the opportunity to implement them and making them available for everyone; on the other hand, it can put human rights at risk due to misuse (e.g. unfairness and bias). Preventing the risks and grasping the opportunities is the objective of human-centered AI.

① Opportunities and risks of AI

- AI opportunities:
  - contribute to education and knowledge;
  - enhance public administration and promote economic development;
  - manage various aspects of society in a better way (e.g. traffic);
  - protect the environment and improve safety;
  - promote participation.

Obs. traditionally, it was assumed that the machine does routine tasks while the human takes decisions, but now that machines have gained a certain level of intelligence they can assist human in taking decisions, or even take decisions by themselves; therefore, there's the need to understand how to integrate human and AI capacities while preserving human initiative and skills

- AI risks:
  - risk for Labour and unemployment, since many workers may become redundant;

·increased economic inequality;
·surveillance systems;
·manipulate people's opinions (e.g. recommender systems);
·polarization through bubbles;
·manipulate information (e.g. fake news).
·Concerning the future of AI, more directions are possible (as opposed to the so-called "technological determinism", namely the idea that there's a single path in technological development)

② How to plan ahead

· Requirements:
·hard sciences like physics to understand how things work;
·technology to understand what is available or possible;
·social sciences to anticipate how technology will impact society;
·normative knowledge to understand what norms and values implement.
· Normative knowledge comprises both general ethical theory (e.g. utilitarianism and deontology) and regulations (e.g. data protection and civil liability).
· Human rights and social values may be the necessary link between ethics and regulations.
· AI for people:
·enabling human self-realisation without devaluing human abilities;
·enhancing human agency without removing human responsibility;
·cultivating social cohesion without eroding human self-determination.
· Trustworthy AI:
·respect for human autonomy;
·prevention of harm;
·fairness;
·explicability.

→ very broad goals, as opposed to human rights which are more lower-level and can thus provide clearer guidance for some issues

③ Human rights

· Human rights are very basic (e.g. thought, assembly, movement, speech, etc.) and are not enough to build a good ICT society (Information and Communication Technologies), but they are a fundamental component of it.
· The same right may be seen differently in different cultures.
· There is a vast disagreement on how to address the conflicts between rights (e.g. privacy vs. freedom of speech), which can be seen positively since different opinions express different valuable positions that allow for debate and improvement.
· According to the Indian philosopher and economist Amartya Sen, human rights are primarily ethical demands which must not be "juridically incarcerated" (since the legal perspective is limited).
· Human rights concern freedoms (opportunities, including liberty and social rights) satisfying some "threshold conditions" of special importance and social influenceability (e.g. tranquility of mind is important but it cannot be a human right, since society cannot deliver it).
· They may lead to:
·imperfect duties (goals that can be achieved as long as there are no other competing goals);
·perfect duties (rights that should never be violated).
· They may be the object of advocacy, political debate and, not always, legal [redacted] enforcement.
· ICT and human rights:
·ICTs can either:
·interfere with human rights;
·contribute to protect and implement human rights;
·provide for the existence of new human rights or add new content for existing rights by [redacted] endowing a [redacted] certain human opportunity with importance and enabling society to realize it (e.g. right to health care or to internet access).
· Human rights should not be seen [redacted] only as an endangered legacy, but also as a blueprint for the [redacted] future.
· Human rights are:
·Ethical rights ⇒ some opportunities are fundamental for humans and achievable by society.
·Political rights ⇒ society should provide ways for the [redacted] fulfillment of these opportunities.

- Legal rights ⇒ such rights should not be violated.

④ Some rights in detail → "Republican" definition of freedom: being free to do thing without being subject to the arbitrary choice of others ("freedom as non-domination")

- Freedom and dignity (1):
  - "All human beings are born free and equal in dignity and rights."
  - On one hand, ICTs have enhanced human freedom by providing an easy access to culture (e.g. internet and Wikipedia) and an easy exchange of ideas (e.g. social networks).
  - On the other hand, ICTs and AI can also limit human freedom (e.g. unemployment, recommender systems manipulating people's opinions).
  - Concerning dignity, it is linked to autonomy and it can be limited by ICTs and AI in the context of autonomous decision making and surveillance systems/social score.
- Right to equality and non-discrimination (7):
  - "All are equal before the law and are entitled without any discrimination to equal protection of the law."
  - "All are entitled to equal protection against any discrimination [...] and against any incitement to such discrimination."
  - Equality of opportunity means being able to make use of one's talent to achieve one's goals.
  - Equality of outcome states that some sort of redistribution is needed s.t. those less likely or less capable can still get a decent life.
  - Equality of opportunity may be promoted by ICTs and AI since they facilitate universal access to culture and information, but may be also threatened by them if they magnify the difference in skill and education (i.e. those more skilled can benefit from technology to be more productive, whereas those less skilled risk to be left aside).
  - To respect the right to nondiscrimination, AI must be fair.
- Right to privacy (12):
  - "No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, not to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks."
  - ICT makes it possible to capture and process a large amount of personal information, and using AI further information can be inferred.
  - Such information can be used for the good of the person (e.g. personalized medicine) but also for his/her bad (e.g. insurance companies or surveillance systems).
  - The right to reputation is also affected by ICT, since the digital image of oneself has become as important as the real one.
  - The right to erasure ("right to be forgotten") concerns the right to obtain the erasure of personal data.
  - The right to identity is the right to have a representation that corresponds to one's reality and not to the way one has been considered.
- Right to life, liberty and security (3):
  - "Everyone has the right to life, liberty and security of person."
  - This right concerns the physical integrity and the right not to be harmed.
  - Such right is challenged by AI in the contexts of autonomous weapons or AVs.
  - We can say that this right goes beyond the body of a person: since, nowadays, devices such as PCs or smartphones can be considered an extension of our minds, even a hacker attack targeting one's smartphone can be considered as a violation of the right to security.
- Right to property (17):
  - "Everyone has the right to own property alone as well as in association with others."
  - The ICT devices we're using may be covered by the right to property whereas the data we're storing in such devices, or especially on the cloud, challenge such right.
  - The right to portability, granted by GDPR, is the right to have one's data exported from the platform where they're stored in a reusable format.
- Freedom of assembly and association (20):
  - "Everyone has the right to freedom of peaceful assembly and association."
  - "No one may be compelled to belong to an association."
  - ICT provides powerful tools to exercise this right (e.g. social networks).
  - On one hand, AI can facilitate surveillance and detection of people participating in unwanted associations by analysing communication (and thus affect negatively such right), on the other hand AI can

enhance the capacity of people to find associative links.

- Right to an effective remedy (8):
  - "Everyone has the right to an effective remedy by the competent national tribunals for acts violating the fundamental rights granted him by the constitution or by [redacted] law."
  - AI could be used to monitor what happens inside a legal system and detect instances of injustice and corruption.
  - On the other hand there are worries concerning the use of AI in the judicial system due to bias (e.g. the COMPAS system): in fact, one can argue that using statistical algorithms to forecast one's recidivism is unfair, since everyone is an individual different from others (even though also humans rely on their generalisation capabilities).
  - In case of smart contracts, an agreement is [redacted] implemented automatically using a blockchain, so if things go wrong there cannot be an effective remedy.
- Right to a hearing (10):
  - "Everyone is entitled in full equality to a fair and public hearing by an independent and impartial tribunal, in the determination of his rights and obligations and of any criminal charge against him."
  - Automated decision making may challenge this right (e.g. when the decision about the denial of a loan, or hiring, are taken by autonomous systems), since such systems are not able to understand eventual complaints.
- Presumption of innocence (11):
  - "Everyone charged with a penal offence has the right to be presumed innocent until proved guilty according to law in a public trial at which he has had all the guarantees necessary for his defence."
  - AI systems predicting the tendency of a person to commit a crime would violate this right
  - This raises a dilemma: is it right to intervene before a crime happens, [redacted] punishing a person who is still innocent, or is it better to intervene after the crime and punishing the guilty?
- Freedom of opinion, expression and information (19):
  - "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."
  - Internet allows to access and exchange information, and AI-aided [redacted] search engines simplify such access.
  - On the other hand, ICT can be used to restrict internet access (e.g. China's Great Firewall) for the purpose of [redacted] political control; AI can be built upon that to analyse communication and detect dissidents.
  - Moreover, there's the problem of [redacted] information manipulation and fake news.
- Right to take part in government (21):
  - "Everyone has the right to take part in the government of his country, directly or through freely chosen representatives."
  - "Everyone has the right to equal access to public service in his country."
  - In some cases (e.g. Cambridge Analytica) ICT and AI were used not to promote people's rational deliberation but to manipulate and influence them.
- Right to social security (22):
  - "Everyone, as a member of society, has the right to social security and is entitled to realization [...] of the economic, social and cultural rights indispensable for his dignity and the free development of his personality."
  - AI could reduce the cost of the management of social services and contribute to an effective, not invasive social security.
  - More in general, AI can contribute to make society more efficient and productive.
- Right to work (23):
  - "Everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment."
  - AI can have negative impact on the right to work, since it can replace human workers in some fields.
  - On the other hand, AI can create new opportunities and jobs.
  - Moreover, ICT systems could be used to [redacted] monitor workers' activity, limiting their dignity and freedom by enforcing constraints.
  - AI could also contribute to reducing the dangers involved in risky activities and improve work

conditions.
- Right to an adequate standard of living (25):
  - "Everyone has the right to a standard of living adequate for the health and well-being of himself and of his family [...] and the right to security in the event of unemployment, sickness, disability, widowhood, old age or other lack of livelihood in circumstances beyond his control."
  - AI can contribute by increasing social productivity s.t., under an appropriate political will, an adequate standard of living can be ensured to everybody.
- Right to education (26):
  - "Everyone has the right to education. Education shall be free, at least in the elementary and fundamental stages."
  - ICT and AI can highly contribute to education, making it available to more and more people around the world and also providing tools for interactive learning.
- Right to culture (26):
  - "Everyone has the right freely to participate in the cultural life of the community, to enjoy the arts and to share in scientific advancement and its benefits."
  - ICT facilitates access to intellectual and artistic [redacted] work as well as the creation of new contents.
  - AI can play a role in this by providing new ways of expressing artistic ideas but also of engaging in science.

Obs. in conclusion, human rights are not only a precious heritage to protect, but also a sort of [redacted] compass to guide us in this ICT/AI revolution towards a human-centered AI.

→ 7. LOGIC PROGRAMMING:

Explainable and Ethical AI: A perspective on Argumentation and Logic Programming (talk):

Autonomous [redacted] agents have been actively developed to be involved in a wide range of fields; [redacted] more complex issues concerning responsibility are becoming more and more critical, in particular [redacted] when the agents face situations involving choices on moral or ethical dimensions.

Since autonomous agents would operate together, in a multi-agent system, two perspectives about machine ethics are possible:
- one stressing individual cognition and behaviour:
  - computation can become the vehicle for the study of morality, [redacted] namely in the computational model and in the design of an agent's knowledge and cognition we can address also morality issues;
  - some Logic programming techniques and extensions can be effective for dealing with ethics design;
  - reasoning features:
    - abduction with integrity constraints;
    - preferences over abductive scenarios; → abductive reasoning allows inferring a fact α as an explanation of another fact β
    - probabilistic reasoning;
    - counterfactual thinking and updating;
    - argumentation.
- One stressing collective morals, and how they emerged.

Logic programming (LP) is considered a good choice to deal with morality since many moral issues and their conceptual viewpoints are close to LP-based representation and reasoning:
- moral permissibility, taking into account different logic models (e.g. double effect, triple effect, contractualism);
- dual process logic model which stresses the interaction between deliberative and reactive processes that are always involved when dealing with moral decisions;
- role of counterfactual thinking in moral reasoning.

① Agents

- Agents are autonomous computational entities, and they encapsulate control along with a criterion to govern it.
- Autonomous agents are interactive, social, proactive and situated.

. They might have goals or tasks, or be reactive.
. They live within a bigger Multi-Agent System (MAS), and interact with other agents through communication actions and with the environment through pragmatical actions.

② Motivation

. Logic-based approaches already play a well-understood role in the engineering of intelligent MAS.
. Declarative, logic-based approaches have the potential to represent an alternative way of delivering symbolic intelligence, complementary to the one pursued by sub-symbolic approaches:
. LP address opaqueness issues and, once suitable integrated with argumentation capabilities, can provide for features like interpretability, observability, accountability and explainability;
. well-founded definition of explanation (abducible, conversational).

Obs. such LP-based reasoning should then be integrated with sub-symbolic approaches.

. LP reasoning features:
  . "abduction" scenario generation and of hypothetical reasoning, including the consideration of counterfactual scenarios about the past;
  . "preferences" enacted for preferring scenarios obtained by abduction;
  . "probabilistic LP" allows abduction to take scenario uncertainty into account;
  . "LP counterfactuals" permit hypothesizing into the past, even taking into account present knowledge;
  . "argumentation" converse, debate and explain;
  . "LP updating" enables updating the knowledge of an agent;
  . "tabling" affords solutions reuse and is employed in joint combination with abduction and updating.
. Advantages of implementing machine ethics with LP:
  . it is a declarative paradigm;
  . it is a tool for knowledge representation;
  . it allows for different forms of reasoning and inference.
  → can lead to properties that are critical in the design of ubiquitous intelligence (in terms of both transparency and ethics)
. Provability is a key feature in the case of trusted and safe systems:
  . correctness, completeness, well-founded extension;
  . ensuring some fundamental computational properties, such as correctness and completeness;
  . extensions can be formalized, well-founded as well, based on recognized theorems.
. Explainability is another important feature:
  . formal methods for argumentation, justification and counterfactual reasoning are often based on LP;
  . explainable systems are capable to engage in dialogues with other actors to communicate its reasoning, explain its choices, or to coordinate in the pursuit of a common goal;
  . other logical forms of explanation can be envisaged via non-monotonic reasoning and argumentation, through a direct extension of the semantics of LP.
. Expressivity and situatedness are two other desirable properties, obtainable with LP:
  . exploit different extensions to inject application-specific expressivity in the program and handle different nuances;
  . make explicit assumptions and exceptions;
  . capture the specificities of the context.
. Hybridization is the last feature we can implement with LP extensions:
  . it allows to integrate different contexts;
  . it allows to represent the heteropeneity of the contexts of intelligent systems (also in relation to application domains) and to customize as needed the symbolic intelligence that is provided while remaining within a well-founded formal framework.
. Why logic for agents:
  . although LP is not an agent programming language with a "theory of agency", it allows to inject logic inference for reasoning, and reasoning for deliberation;
  . it allows to explicitly define belief and represent goals for agent-oriented operations;
  . it could be used to build cognitional artefacts.

Obs. actually we can build more specific agent language leveraging LP.

③ Essentials of LP and Prolog

- Fundamenta features:
  - Terms ⇒ computing takes place over the domain of all terms defined over a "universal" alphabet
  - MGU ⇒ values are assigned to variables by means of automatically-generated substitutions, called "most general unifiers", which may contain the so-called "logical variables".
  - Backtracking ⇒ control is provided by a single mechanism, "automatic backtracking".
- Let A be an alphabet of a language L, which is a countable disjoint set of constants, function symbols, and predicate symbols.
- An alphabet is assumed to contain a countable set of variable symbols.
- A term over A is defined recursively as either a variable, a constant or an expression of the form $f(t_1,\dots,t_n)$ where f is a function symbol of A and t_i are terms.
- An atom over A is an expression of the form $p(t_1,\dots,t_n)$ where p is a predicate symbol of A and t_i are terms.
- p/n denote the predicate symbol p having arity n.
- A Literal is either an atom a or its negation ¬a.
- A term (respectively, atom and Literal) is "ground" if it does not contain variables.
- The set of all ground terms (respectively, ground atoms) of A is called the Herbrand universe (respectively, Herbrand base) of A.
- In Prolog:
  - Terms are built recursively out of functors and variables as in LP:
    - Variables ⇒ alphanumeric strings starting with either an uppercase letter or an underscore
      - ↳ underscore alone is the anonymous variable
      - ↳ underscore followed by a string is a normal variable during resolution, but it does not need to be exposed in the computed substitution.
    - Functors ⇒ alphanumeric strings starting with a lowercase letter (both proper functors and constants).

E.g. term, Var, f(X), p(Y, f(a)) are Prolog terms
term, var, f(a), p(x, y) are Prolog ground terms

  - Atoms are built applying predicates to terms as in LP:
    - Predicates ⇒ alphanumeric strings starting with a lowercase letter (same as functors).
  - Clauses:
    - Clause ⇒ a Horn clause of the form $A :- B_1,\dots,B_n.$
      - ↳ $A, B_1,\dots,B_n$ are Prolog atoms
      - ↳ A is the head of the clause
      - ↳ $B_1,\dots,B_n$ is the body of the clause
      - ↳ :- denotes logic implication
      - ↳ . is the terminator
    - fact ⇒ a clause with no body, e.g. A. (n=0)
    - rule ⇒ a clause with at least one atom in the body, e.g. $A :- B_1,\dots B_n.$ $(n>0)$
    - goal ⇒ a clause with no head and at least one atom in the body, e.g. $:- B_1,\dots,B_n.$ $(n>0)$ (usually written as $?:- B_1,\dots,B_n.$)
  - A Prolog program is a sequence of Prolog clauses interpreted as a conjunction of clauses.
  - A logic theory is a theory made of Horn clauses.
- Prolog execution:
  - Aim:
    - given a Prolog program P and the goal $?:- p(t_1,\dots,t_m)$, also called query;
    - if $X_1,\dots,X_n$ are the variables in terms $t_1,\dots,t_m$;
    - the aim of the Prolog computation is to query P and find whether there are some values for $X_1,\dots,X_n$ that make $p(t_1,\dots,t_m)$ true, namely it's to find a substitution $\sigma = X_1/s_1,\dots,X_n/s_n$ s.t. $P \models p(t_1,\dots,t_m)\sigma$.
  - Search strategy:
    - Prolog adopts SLD resolution;
    - it applies SLD in a strictly linear fashion (goals are replaced left-to-right, clauses are considered in top-to-bottom order, subgoals are considered immediately once set up) ⇒ depth-first

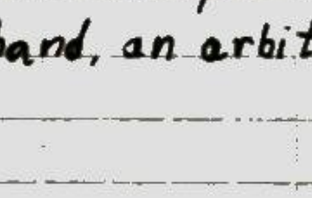
- Backtracking:
  - to achieve completeness, Prolog saves choice points for any alternative still to be explored;
  - in case of failure, it goes back to the nearest available choice point (automatic backtracking).

- Abduction extension:
  - The notion of abduction is characterized as "a step of adopting a hypothesis as being suggested by the facts".
  - It's a type of reasoning in which one chooses from the available hypotheses those that best explain the observed evidence.
  - It is implemented by extending LP with abductive hypotheses, called abducibles.
  - Abductive logic programs have three components $\langle P, AB, IC \rangle$ where:
    - P is the Logic program;
    - AB is a set of predicate names, called abducible predicates;
    - IC is a set of first-order classical formulae.
      - ↳ Integrity Constraint

E.g.
Grass is wet if it rained. → rules
Grass is wet if the sprinkler was on. → rules
The sun was shining. → fact
IC: false if it rained and the sun was shining. → IC

The observation that "the grass is wet" has two potential explanations, "it rained" and "the sprinkler was on", but only the latter satisfies the integrity constraint.

- Abstract argumentation:
  - An argumentation system consists of a couple (A,R), where A is a set of elements (arguments) and R is a binary relation representing attack relation between arguments:
    - it's represented by a directed graph;
    - each node represents an argument;
    - each arc denotes an attack by one argument on another.



  - The graph is analysed to determine which arguments are acceptable according to some general criteria; (so-called "acceptability criteria").
  - The process of deciding which arguments should be accepted or discarded is called argument evaluation
  - Common approaches:
    - Extension-based ⇒ semantics specification concerns the generation of a set of extensions ("collective acceptable" arguments) from an argumentation framework.
      - ↳ determine conflict-free sets
      - ↳ determine extensions (naive, admissible, preferred, complete, stable, etc.)
    - Labelling-based ⇒ semantics specification concerns the generation of a set of labellings (e.g. possible alternative states of an argument) from an argumentation framework.

Obs. any extension-based argument can equivalently expressed in a simple labelling-based argument adopting a set of two labels (e.g. $L = \{in, out\}$); on the other hand, an arbitrary labelling cannot in general be formulated in terms of extensions.

- Extension-based approaches:
  - Four traditional semantics (Dung's paper):
    - complete ⇒ a set able to defend itself and including all arguments it defends;
    - grounded ⇒ includes those and only those arguments whose defense is rooted in initial arguments (also called strong defense);
    - stable ⇒ it attacks all arguments not included in it;
    - preferred ⇒ the aggressive requirement that an extension must attack anything outside it may be relaxed by requiring that an extension is as large as possible and able to defend itself from attacks.
  - Subsequent proposals introduced by various authors in the literature, often to overcome some limitation or improve some undesired behaviour of a traditional approach: stage, semi-stable, ideal, CF2, prudent semantics.

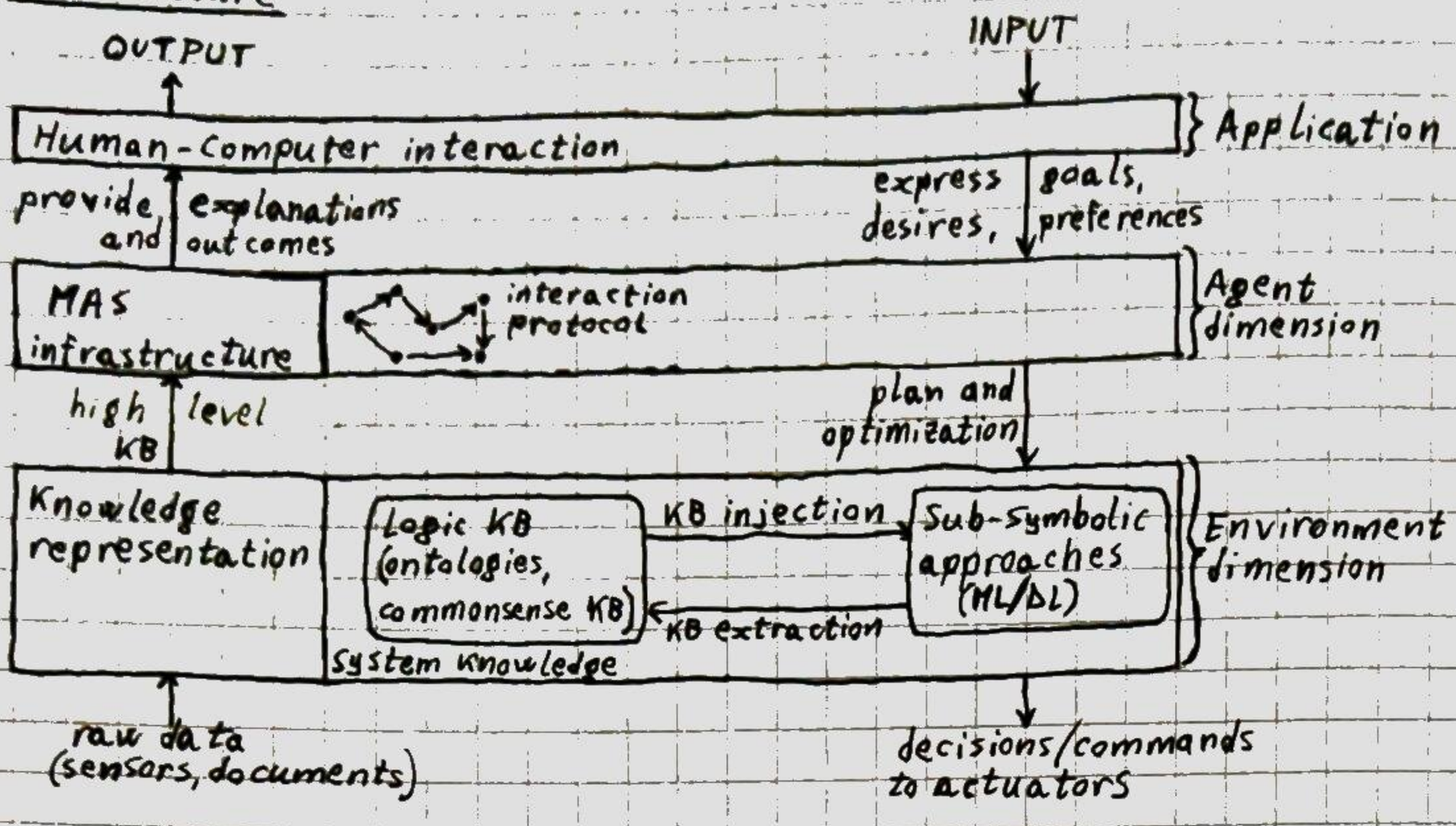
④ LP approach to Ethics

- Abduction:
  - it enables the generation of plausible scenarios under certain conditions, and hypothetical reasoning including the consideration of counterfactual scenarios about the past;
  - counterfactual reasoning suggests thoughts about what might have been, what might have happened if any event had been different in the past;
  - it provides hints about the future by allowing for the comparison of different alternatives inferred from the changes in the past;
  - it provides justification of why different alternatives would have been worse;
  - it provides integrity constraints, which exclude abducibles that have been ruled out a priori
  - a posteriori preferences are appropriate for capturing utilitarian judgements that favours welfare-maximising behaviours;
  - it combines a priori integrity constraints and a posteriori preferences, resulting in a model which reflects the dual process of intuition and reflection;
  - a priori integrity constraints are a mechanism to generate immediate responses in deontological judgement;
  - reasoning with a posteriori preferences can be viewed as a form of controlled cognitive processes in utilitarian judgement: after excluding those abducible that have been ruled out a priori by the integrity constraints, the consequences of the considered abducibles have first to be computed, and only then are evaluated to prefer the solution affording the greater good
- Probabilistic logic programming (PLP):
  - it enriches symbolic reasoning with degrees of uncertainty;
  - it allows abduction to take scenario uncertainty measures into account;
  - it accounts for diverse types of uncertainty, in particular uncertainty on the credibility of the premises, uncertainty about which arguments to consider, and uncertainty on the acceptance status of arguments or statements;
  - one of the key factors that allow a system to fully manage the formulation of well-founded reasoning on which scenario to prefer and which suggestions to provide as outcomes.
- Argumentation:
  - it enables system actors to talk and discuss in order to explain and justify judgements and choices, and to reach agreements;
  - despite the long history of research in argumentation and the many fundamental results achieved, much effort is still needed to effectively exploit argumentation in distributed and open environments.
- Sample scenarios:
  - Princess saviour moral robot:
    - a princess is held in a castle;
    - a robot must rescue her;
    - the path to the castle is blocked by a river, crossed by two rivers;
    - each bridge is guarded, so the robot must defeat one guard to proceed;
    - prospective reasoning is the combination of pre-preference hypothetical scenario generation of the future plus post-preference choices taking into account the imagined consequences of each preferred scenario;
    - by reasoning backwards from the goal, the agent generates three possible hypothetical scenarios for actions, either it crosses one of the two bridges or it does not cross the river at all (thus negating the satisfaction of the rescue goal);
    - in order to derive the consequences for each scenario, the agent has to reason forward from each available hypotheses;
    - as soon as these consequences are known, meta-reasoning techniques can be applied to prefer amongst partial scenarios;
    - supposing that one guard is weaker than the other, it will be more convenient for the robot to fight the weaker guard, since the chances to survive are higher (the imperative to survive will be encoded in a utilitarian rule, together with the goal);
    - another "knight rule" will state that the agent must save the princess no matter what (not a utilitarian behaviour);
    - in case of no morality rules, both rules are retracted and the robot does not save the princess

since it has no intent to do so;
- in order to maximise its survival chance in saving the princess, the robot updates itself with utilitarian moral and decides to fight the weaker guard;
- however, supposing that the weaker guard is a human while the stronger one is not, and supposing that the princess follows a deontological ethics and argues that the robot must not kill the human guard, then the new deontological rule would conflict with its utilitarian rule, resulting in the robot leaving the mission;
- by imposing the "Knight rule", the robot fights the stronger guard, is defeated and thus fails the mission;
- the justifications of the robot's actions are always clear;
- the argumentation caused by the conflicting rules is carried out until an agreement is reached.

- Autonomous cars:
  - a road is equipped with two traffic lights, one for the vehicles and one for the pedestrians;
  - the goal of the system is to autonomously manage intersections accordingly to traffic light indications;
  - during emergencies, authorized vehicles can ignore traffic light prescriptions, while other vehicles must leave the way clear;
  - two rules will encode fundamental constraints "if the traffic light is red, the road users have to stop" and "otherwise, they can proceed";
  - a rule will encode the emergency situation, and another one will give authorized vehicles the permission to proceed in emergency situations even if the light is red;
  - another rule will impose users the obligation to stop if aware of another vehicle in emergency state;
  - two preferences will assign a higher priority to emergency situations over ordinary ones;
  - supposing that there are three users on the road (a car, an ambulance and a pedestrian), that the ambulance has it's acustic and light indicators on (emergency situation), and that the traffic light is red for both ambulance and car, and green for the pedestrian, then the pedestrian has to stop since the emergency situation has higher priority over the light being green for him, while the ambulance can go on even though the light is red for it;
  - supposing that the pedestrian, despite his obligation to stop, continues the crossing and gets hit by the ambulance, which fails to see him, then we must find the responsibilities of the parties involved in the accident;
  - supposing that the case is under the Italian jurisdiction, which states that responsibility in an accident is based on the concept of carefulness, both the ambulance driver and the pedestrian have to prove that they were careful and acted according to the law;
  - the ambulance driver's action of not stopping at the red light is legitimate due to the emergency, and for the same reason the pedestrian's action is not legitimate;
  - however, supposing that one witness claims that the ambulance was proceeding at proper speed whereas another witness claims that it was proceeding at high speed, then there is an uncertainty on the ambulance driver's carefulness, and such uncertainty is considered as a failure to meet the burden of persuasion, therefore both are considered responsible;
  - now let's suppose that the ambulance driver declares that he tried to stop the ambulance but the brake didn't work, and that a mechanic confirms the issue;
  - then, the ambulance's manufacturer is called to prove that the ambulance was not defective when delivered (i.e. the burden of proof on the adequacy of the vehicle is on the manufacturer);
  - the discovery of a defect in the ambulance would lead to the discarding of the driver's responsibility, and if the manufacturer fails to meet his burden it would share the responsibilities of the accident;
  - even though the manufactures declares that every vehicle is deeply tested before the delivery (including the one at hand), he fails to provide the related documentation;
  - in conclusion, the ambulance driver is free from every responsibility in the accident since his prudence is correctly proven, whereas the manufacturer is found responsible.

⑤ Architecture



- Knowledge is built by synergically exploiting both logic and sub-symbolic approaches.
- Logic allows the injection of ethical behaviour in the system.
- The agent exploits the knowledge built to interact in the system (it also complies to institutional norms).
- The agent can then provide explanations to the user, as well as receive goals or preferences as input which are converted to plans and finally to actual commands to actuators.
- The tuProlog system makes two different, complementary technologies available to build MAS abstractions:
  - Kotlin/Java to implement deterministic, object-oriented parts of an abstraction;
  - Prolog to create non-deterministic, logic-based parts of an abstraction.

⑥ Limits

- The tuProlog agent systems have some problems:
  - they are closed systems, meaning that no new agent apart from the ones originally envisioned by the designer can enter the system;
  - the expressive power of abstractions available in tuProlog is not [redacted] enough to capture the elements of MAS models:
    - Prolog engines alone do not lead to the creation of robust MAS, not even single agents;
    - Prolog engines are the most high-level abstraction in the system;
  - basic communication and coordination infrastructures need to be implemented from scratch;
  - building such infrastructures would possibly require a huge ad-hoc effort.
- To leverage MAS, other kinds of programmable supports are needed.
- The tuProlog engine can be the basic brick for those kinds of fundamental layers:
  - coordination infrastructures based on a declarative, logic-based programming model;
  - new [redacted] logic languages providing more powerful abstractions as first class entities;
  - pattern-based matching for communication facilities.

→ 8. MODELLING NORMS:

Modelling norms (talk):

The development of a computable law consists in two steps:

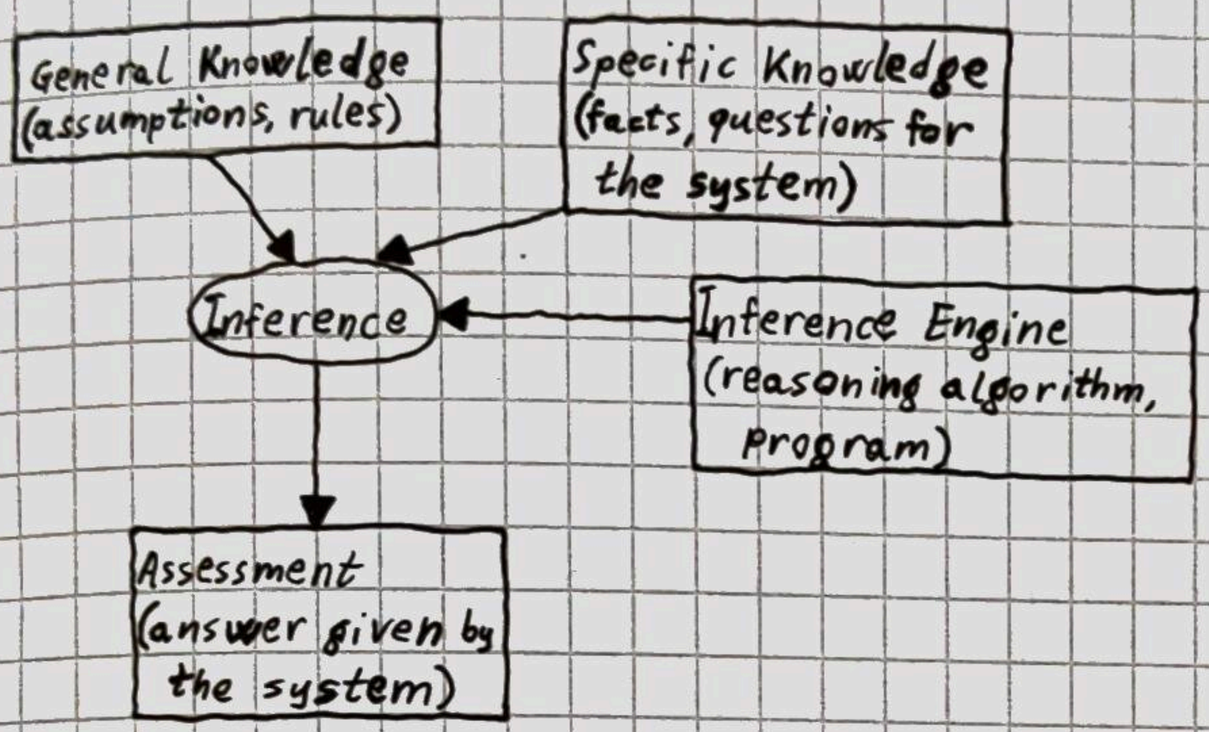
1. Modelling and formalization of the law:
   - Input ⇒ sources, cases, concepts, doctrines
   - Output ⇒ computable models (knowledge base)
   - Process ⇒ Logic programming/Knowledge representation
2. Feeding algorithms with computable models:
   - Input ⇒ computable models of the law
   - Output ⇒ answers, legal qualifications, support to decision making
   - Process ⇒ forward and backward rule chaining, deduction, defeasible reasoning

① Knowledge representation

- "Knowledge representation is the application of logic and ontology to the task of constructing computable models for some domain" (Sowa, 2000).
- Logic provides the formal structure and rules of inference.
- Ontology defines the kinds of things that exist in the application domain, and their interrelationship.
- Computable models implement logic and ontology into computer systems and applications.
- Declarative programming languages are well suited for this task:
  - the program consists of logical statements, expressing the knowledge about the domain in terms of known facts and relationships;
  - the program executes by searching for proofs of the statements.

② Legal applications

- Legal rule-based systems allow for transfering general knowledge about the legal domain to computer systems.



- By 1980s, a number of researchers had implemented working systems based on manually-created logical representations of rules (e.g. Sergot et al. in 1986 realised a computable model of the British Nationality Act).
- Rule-based systems are used in the legal domain for legal analysis and automated legal assessment, and have also many applications in public administration and in business.
- In legal domain, a typical reasoning scheme is the application of rules since, in fact, legal rules may be seen as conditional if-then statements, linking an antecedent to a consequent s.t. from the former is possible to infer the latter.
- This corresponds to the idea that legal rules usually connect a set of abstract provision of facts to a legal effect.

E.g. if a person Y commits the crime X, then Y shall be punished with sanction Z.

- Formal and computable languages: Prolog, RuleML, and various commercial solutions.
- In general we'll have a premise which comprises a rule and a fact, and a conclusion representing the legal effect.
- Issues with legal knowledge representation:
  - laws' ambiguity, vagueness (open texture), and density of meaning;
  - logic representation's rigidity, in contrast with the laws which are flexible (to some extent) to capture the real world's complexity;
  - since laws are about how the world ought to be, there's the need to represent with logic deontic positions as obbligation or permission;
  - how to enable temporal reasoning to represent law as a dynamic system;
  - how to deal with conflicting legal rules and/or rules that can be excluded from being applicable by other rules (i.e. defeasible reasoning);
  - how to manage reification, whenever rules representing legal norms need to be treated as objects with properties by other rules;
  - how to mantain isomorphism between source text and logic representation, needed to guarantee explainability;

- bottleneck due to knowledge elicitation, representation and update.

E.g. Article 17 of Montreal Convention for Aviation:
"The carrier is liable for damage sustained in case of death or bodily injury of a passenger upon condition only that the accident which caused the death or injury took place on board of the aircraft or in the course of any of the operations of embarking or disembarking."

⇓

if Carrier(x) ∧ Accident(y) ∧ Passenger(p) ∧ Caused(y,f) ∧ DeathOrInjury(f) ∧ Object(p,f) ∧ [TookPlaceOnBoard(y) ∨ InCourseOfEmbarking(y) ∨ InCourseOfDisembarking(y)] ∧ ArisingFrom(d,f) ∧ Damage(d)
then LiableToFor(x,p,d)

E.g. Article 615/ter of Italian criminal code about unauthorized access to a computer system:
"Whoever enters a computer or telecommunication system which is protected by security measures or remains in such system against the will of the person who is entitled to exclude him shall be punished with detention up to three years."

⇓

[if [a: the individual enters the computer or telecommunication system]
and [b: the computer or telecommunication system is protected by security means]
or [c: the individual remains in the computer or telecommunication system]
and [d: there is the contrary will of the person who is entitled to exclude the individual]
then [e: the individual shall be punished with detention up to three years]

→ ambiguity about the ~~order~~ logic order

- Vagueness or open texture: "All rules involve recognizing or classifying particular cases as instances of general terms, and in the case of everything which we are prepared to call a rule it is possible to distinguish clear central cases, where it certainly applies and others where there are reasons for both asserting and denying that it applies. Nothing can eliminate this duality of a core of certainty and a penumbra of doubt when we are engaged in bringing particular situations under general rules. This imparts to all rules a fringe of vagueness or "open texture" (Hart, The Concept of Law).

Obs. cf. Sergot et al. paper "The British Nationality Act as a logic program"

E.g. Article 91 of 5 February 1992 about Italian nationality:
"1. The following shall be citizens by birth:
a) any person whose father or mother are citizens;
b) any person who was born in the territory of the Republic, either where both parents are unknown or stateless, or where he or she does not acquire his or her parents' citizenship according to the law of the state to which the latter belong.
2. Any person who is found in the territory of the Republic, whose parents are unknown, shall be deemed a citizen by birth, where their possession of any other citizenship cannot be proven."

⇓

citizen(A) :- (father(B,A); mother(B,A)), citizen(B).
citizen(A) :- born_in_republic(A), (father(unknown,A), mother(unknown,B);
father(X,A), mother(Y,A), stateless(X), stateless(Y);
father(X,A), mother(Y,A), citizen_of(X,R), citizen_of(Y,S), citizenship_not_inheritable(R), citizenship_not_inheritable(S)).

③ Oracle Policy Automation (OPA)

- Suite of tools that supports the creation and deployment of rule-based knowledge systems, helping the rapid writing of rules with an integrated rule editor, validation/mass testing tools, easy development and customization of user interfaces.
- Rules are written in a customized MS Word environment using a quasi-natural language.
- A linguistic component, called parser, analyses the syntactic structure of phrases in order to identify their logical components.

- Rules are then translated into an XML-based format, used by the Inference Engine.
- The linguistic component automatically prepare questions and explanations for the user interface.

Interlex (talK):

It's a project meant to deal with issues of international private law, and it's implemented in Prolog. Extensions, as Constraint Logic Programming (CLP), Constraint Handling Rules (CHR) and Hypothetical/Temporal Reasoning (Sciff), together with meta-interpreters, were used to extend Prolog's functionality.

E.g. Article 4.1:
"Subject to this regulation, persons domiciled in a Member State shall, whatever their nationality, be sued in the courts of that Member State".

⇓

```
hasGeneralJurisdiction(CountryCourt, ClaimId, brusselsRegulation):-
    personRole(PersonId, ClaimId, defendant),
    personDomicile(PersonId, Country, Court),
    memberState(Country).
```

To handle more complex articles which are divided in multiple parts, one may develop a generic rule for the main article and a rule for each subpart.

① Ontologies

- To model complex sets of rules, one may define an ontology:
  - define who the actors are;
  - define their properties.

E.g. In this ontology, the actor "Person" has the following properties:
- Nature (Legal/Natural);
- Role (Plaintiff/Defendant/Third party) + Claim Id;
- Type (Consumer/Business/Employer/Employee/Insurer/Trust);
- Work (Country);
- Activity In (Country);
- Domicile (Country);
- Establishment (Country);
- PersonId.

The actor "Claim" has the following properties:
- Matter (Civil/Commercial);
- Grade (First/Appeal);
- Type (Original/Counter/Incidental);
- Object (Contract/Tort/Ownership/Rights in Rem/Liability/Trust);
- ClaimId.

② SWI-Prolog

- Advantages:
  - rules are compact and readable;
  - the logical structure of rules matches the legal text;
  - exceptions can be easily introduced;
  - the closed-world assumption is implemented in the system;
  - developments are possible by using available tools for temporal/abductive and hypothetical reasoning (particularly Sciff);
  - the programming environment provides resources for interfaces (forms, queries and printouts) and explanations (through meta-interpretation or other techniques).
- Issues:
  - priorities between rules are not natively modelled (but can be captured using negation-as-failure).

→ 9. DATA PROTECTION:

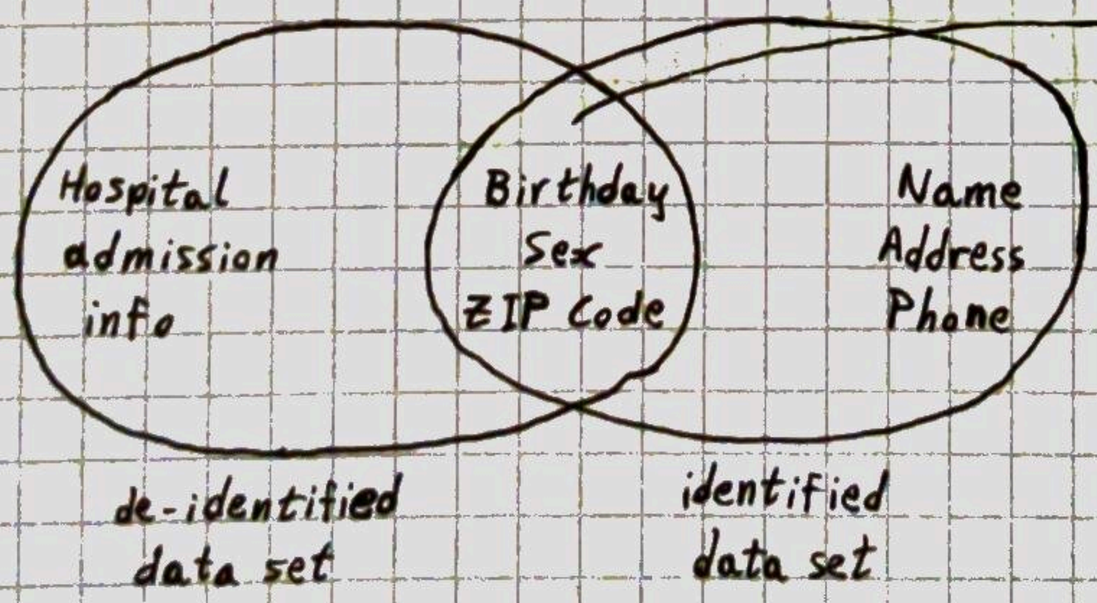
AI in the GDPR:

The General Data Protection Regulation (GDPR) is the main instrument we have today to regulate computing and personal data management. It's more focused on the issues emerging from the internet rather than to AI, but many AI provisions are relevant to GDPR.

① GDPR and Personal Data

- The concept of personal data plays a key role in the GDPR, characterising the material scope of the regulation.
- The provisions in GDPR only concern personal data, to the exclusion of information not related to humans / particular identifiable individuals, and anonymized data.
- In article 4, "personal data" means any information relating to an identified or identifiable natural person ("data subject"), who is one that can be identified directly or indirectly, in particular by reference to an identifier (such as a name, an identification number, location data, etc.).
- Namely, when I have a piece of information which is accompanied by other data referring to the same individual in such a way that I can identify the individual concerned, then such data is personal data.
- AI's powerful inference capabilities may [redacted] change the notion of personal data, as it may extrapolate information concerning identifiable natural people from seemingly anonymous data ("re-personalization" of anonymous data and reidentification of the individuals); moreover, it may infer further personal information from already-available personal data.
- Reidentification:
  - AI, and more generally methods for computational statistics, increases the identifiability of apparently anonymous data, since they enable non-identified data, including data having been anonymised or pseudonymised, to be connected to the individuals concerned.
  - Reidentification of data subjects is usually based on statistical correlations between non-identified data and personal data concerning the same individuals.

E.G. Connection between identified and de-identified data



- Reidentification is a kind of inference that consists in linking personal identifiers to previously non-identified data.
- Thanks to AI and Big Data, the identifiability of data subjects [redacted] has vastly increased.
- This problem can be addressed in two ways, neither of which is failproof:
  - [redacted] ensure that data is de-identified in ways that make it more difficult to re-identify the data subject;
  - implement security processes and measures for the release of data that contribute to this outcome.
- Inference of new information:
  - AI systems may infer new information about data subjects by applying algorithmic models to their personal data.
  - The key issue is whether the inferred information should be considered as new personal data, distinct from the original data from which it has been inferred.
  - If the inferred information counts as new personal data, then automated inferences would

trigger all the consequences that the processing of personal data entails according to the GDPR (i.e., the need of a legal basis, the conditions for processing sensitive data, the data subject's rights, etc.).
- one solution may be to ask for permission beforehand

② Profiling

- "Profiling" means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person's performance at work, economic situation, health, personal preferences, etc.
- Profiling aims at classifying people into categories of groups sharing the features being inferred.
- AI and Big Data, in combination with the availability of extensive computer resources, have vastly increased the opportunities for profiling.
- Assume that a classifier has been trained on a vast set of past examples, which link certain features of individuals (the predictors) to another feature of the same individuals (the target).
- Through the training, the system has learned an algorithmic model that can be applied to new cases, namely given predictor values concerning new individuals it infers a corresponding target value for that individual (ie. a new data item concerning him/her).
- A learned correlation may also concern a person's propensity to respond in certain ways to certain stimuli, enabling the transition from prediction to behaviour modification (both legitimate influence and illegal or unethical manipulation).
- It's necessary to distinguish the general correlations that are captured by the learnt algorithmic model, and the results of applying that model to the description of a particular individual.

E.g. Loan applications:
- consider a ML system that has learnt a model from a training set of previous loan applications and outcomes;
- the system's training set consists of personal data, but the learnt algorithmic model no longer contains [redacted] them, since it links any possible combination of possible input values (predictors) to a corresponding likelihood of default on the loan (target);
- the correlations embedded in the algorithmic model are not personal data, since they apply to all individuals sharing similar characteristics [redacted] (we can possibly view them as group data, concerning the set of such individuals);
- assuming that the algorithmic model is applied to input data consisting of a new applicant (in order to determine that applicant's risk of default), both the description of the applicant and the default risk attributed to him/her by the model represent personal data (the former being collected, the latter being inferred).

- Since inferred data concerning individuals also are personal data under GDPR (at least when they're used to derive conclusions that are or may be acted upon), data protection rights should in principle also apply on inferred data (though concurrent remedies and interests have to be taken into account).
- According to the Article29 Working Party, in case of automated inferences (i.e. profiling) data subjects have a right to access both the personal data used as input for the inference, and the personal data obtained as (final or intermediate) inferred output.
- On the contrary, the right to rectification only applies to a limited extent:
  - when the data are processed by a public authority, it should be considered whether review procedures already exist;
  - in case of processing by private controllers, the right to rectify the data should be balanced with the respect for autonomy of private assessments and decisions.
- According to Article 29 Working Party, data subjects have a right to rectification of inferred information not only when the inferred information is "verifiable" (i.e. it's correctness can be objectively determined), but also when it is the outcome of unverifiable or probabilistic inferences (in the latter case, rectification may be needed not only when the statistical inference was mistaken, but also when the data subject provides specific additional supporting a different, more specific statistical conclusion).

- Right to "reasonable inference":
  - some lawyers have been arguing that automatic inference should respect some [redacted] standards;
  - in particular, inference should not be based on discriminating features such as race or sexual orientation;
  - the data that is being inferred should be relevant for the decision purpose;
  - finally, there should be reliability both in the training set and in the statistical methods to process data.

③ Consent

- According to Article 4 of GDPR, consent should be freely given, specific, informed and unambiguous, and be [redacted] expressed through a clear affirmative action.
- Consent means an indication of a data subject's wishes by which he/she signifies agreement to the processing of personal data.
- An issue of whether consent is really free is that in many cases refusing to give consent [redacted] results in not being able to use a service.
- Two main criticisms:
  1. Consent is most often meaningless, since it is usually not based on real knowledge of the processing at stake, nor on a real opportunity to choose:
     - today's processing of personal data is so complex that most data subjects do not have the skills to understand them and anticipate the involved risks;
     - even if data subjects possessed such skills, they would [redacted] not have time and energy to go through the details of each privacy policy;
     - a refusal to consent may imply the impossibility to use services that are important or necessary to the data subject.
  2. When targeted on specific purposes, consent does not include future, often unknown uses of the data (even when socially beneficial).

④ AI and Big Data

- According to GDPR, consent must be specific to a particular activity (e.g. to use a service).
- However, the collected data are usually used for targeted advertising (which should need a separate consent).
- Moreover, there's the issue of freedom to refuse to give consent while managing to use the service.
- Namely, AI and Big Data raise three key issues:
  1. Specificity.
  2. Granularity.
  3. Freedom.

4.1 Specificity of consent

- The data subject must know the purpose for which the data are going to be processed when asked for the consent.
- Data may be further processed for other purposes if these are not incompatible.
- The requirement of specificity is attenuated for scientific research, as stated in Recital 33, which allows consent to be given not only for specific research projects, but also for areas of scientific research (without specifying the particular purpose).

4.2 Granularity of consent

- The idea behind granularity is "separate consents for separate activities".
- Two implications for AI:
  - the data subject should not be required to jointly consent to essentially different kinds of AI-based processing;
  - the use of a service should not, in principle, be dependent on an agreement to be subject to profiling practices (namely, [redacted] consent to profiling must be separate from access to the

service).

4.3 Freedom of consent

- This issue concerns whether consent is actually free when the user is interested in using a service which is exercised in a condition of monopoly.
- In many cases, to use a service there is no alternative than to deliver personal data.
- The European Council decided that it's permissible to ask for the consent as a necessary condition for the use of a service (there may be doubts when the service is exercised in a condition of monopoly).

⑤ AI and data protection principles

- AI and Big data challenge key data protection principles:
  1. Fairness and transparency.
  2. Purpose limitation.
  3. Data minimisation.
  4. Accuracy.
  5. Storage limitation.

  } Article 5

- Transparency:
  - it means that users (namely people interacting with a system) should know how their data are being processed;
  - this idea is related, but distinct, from the idea of explainable AI (which involves building a "scientific" model of the functioning of an AI system, rather than providing sufficient information to lay people).
- Fairness:
  - users should not be deceived concerning the processing of their data;
  - there is a discussion within the data protection commission about what is called "dark patterns", namely ways in which websites' UIs are designed that trick users into accepting data processing (mainly by presenting visual elements in ambiguous ways);
  - informational fairness is also linked to accountability, since it presumes that the information to be provided makes it possible to check for compliance;
  - there are specific AI-issues because of the complexity of the processing involved in AI-applications, the uncertainty of its outcome and the multiplicity of its purposes
  - the Recital 71 points to a different dimension of fairness called substantive fairness, which concerns the use of appropriate mathematical or statistical procedures for profiling, and the implementation of technical and organisational measures to ensure that there are no errors, that data are secured and that decisions are not discriminatory.
- Purpose limitation:
  - according to purpose limitation, data should be collected and processed only for a purpose that is specified, explicit and legitimate;
  - the idea behind AI and Big Data, conversely, is that once data has been collected it can be used for new purposes (since data can be used to discover hidden patterns, previously unknown);
  - to reconcile purpose limitation and repurposing, some reuse of the data is acceptable as long as it's not incompatible with the original purpose.

→ 10. FAIRNESS IN AUTOMATED DECISIONS:

Fairness in Algorithmic Decision Making (talk):

The combination of AI and Big Data enable automated decision making even in domains requiring complex choices that can be based on several factors and non-predefined criteria. A wide debate has taken place on prospects and risks of algorithmic assessments and decisions concerning individuals.

① AI assessment capabilities

- In many domains automated predictions and decisions are not only cheaper but also more

precise and impartial than human ones.
- AI can avoid typical fallacies of human psychology (such as overconfidence, loss aversion, confirmation bias, etc.) and the human inability to process statistical data, as well as prejudices (e.g. ethnicity, gender or social background).
- In fact, algorithmic systems have often performed better, according to usual standards, than human experts in domains such as investments, recruitments, creditworthiness and judicial matters.
- However, algorithmic decisions may be mistaken or discriminatory:
  - only in rare cases will algorithms engage in explicit unlawful discrimination, so-called disparate treatment, basing their outcomes on prohibited features (predictors) such as race, ethnicity or gender;
  - more often a system's outcome will be discriminatory due to its disparate impact, since it disproportionately affects certain groups without an acceptable rationale (as in the case of COMPAS).

② Main causes of discrimination

- Systems based on supervised learning may be trained on past human judgements and may therefore reproduce the strengths and weaknesses of the humans who made these judgements, including their propensities to error and prejudice.

E.g. A recruitment system trained on the past hiring decisions will learn to emulate the managers' assessment of the suitability of candidates, rather than to directly predict an applicant's performance at work: thus, if past decisions were influenced by prejudice, the system will reproduce the same logic.

- Prejudice baked into training sets may persist even if the inputs (the predictors) to automated systems do not include forbidden discriminatory features (e.g. ethnicity or gender) ⇒ this may happen whenever a correlation exists between discriminatory features and some predictors.

E.g. Assume that a prejudiced human resources manager did not hire applicants from a certain ethnic background, and that people with that background mostly live in certain neighbourhoods: a training set of decisions by that manager will teach the system not to select people from those neighbourhoods, which would entail continuing to reject applications from the discriminated-against ethnicity.

- In other cases, a training set may be biased against a certain group, since the achievement of the outcome being predicted (e.g. job performance) is approximated through a proxy that has a disparate impact on that group

E.g. Assume that the future performance of employees is only measured by the number of hours worked in the office: this outcome criterion would lead to past hiring of women (who usually work for fewer hours than men, having to cope with family burdens) being considered less successful than the hiring of men, and thus the system will predict a poorer performance of female applicants.

- In yet other cases, mistakes and discriminations may pertain to the machine-learning system's biases embedded in the predictors:
  - a system may perform unfairly since it uses a favourable predictor that only applies to members of a certain group (e.g. the fact of having attended a socially selective high-education institution);
  - unfairness may also result from taking biased human judgements as predictors (e.g. recommendation letters).
- Finally, unfairness may derive from a dataset that does not reflect the statistical composition of the population:
  - members of a certain group may also suffer prejudice when that group is only represented by a very small subset of the training set;

. this would reduce the accuracy of predictions for such group.

E.g. Assume that in applications for bail or parole, previous criminal record plays a role, and that members of a certain group are subject to stricter controls s.t. their criminal activity is more often detected and acted upon: this would entail that members of such group will generally receive a less favourable assessment than members of other groups having behaved in the same ways.

③ Challenging unfairness in automated decision making

. It has been observed that it is difficult to challenge the unfairness of automated decision making.
. Challenges raised by the individuals concerned, even when justified, may be disregarded or rejected because they interfere with the system's operations, giving rise to additional costs and uncertainties.
. Predictions of ML systems are based on statistical correlations against which it may be difficult to argue on the basis of individual circumstances, even when exception would be justified.
. On one hand, some experts as O'Neil argue that using statistical algorithms for assessment is dangerous, since the statistical score produced by the system can "turn someone's life upside down" (he calls them "Weapons of Math Destruction").
. On the other hand, otherexperts as Sunstein say that, with appropriate requirements in place, algorithms would make it possible to more easily examine and interrogate the entire decision process, thereby making it far easier to know whether discrimination has occurred, by forcing a new level of specificity, the use of algorithms also highlights certain trade-offs among competing values ("algorithms [...] have the potential to be a positive force for equity").
. The criticisms by O'Neil have been countered by observing that algorithmic systems, even when based on ML, are more controllable than human decision-makers, their faults can be identified with precision, and they can be improved and engineered to prevent unfair outcomes.
. The alternative to automated decision making is not perfect decisions made by humans with all their flaws; a biased algorithmic system can still be fairer than an even more biased human decision-maker.
. In many cases, the best solution is integrating human and automated judgements, by enabling the affected individuals to request a human review of an automated decision as well as by favouring transparency and developing methods which allow human experts to analyse and review automated decision making.
. AI systems have demonstrated an ability to successfully act also in domains traditionally entrusted to human expertise (e.g. medical diagnosis, financial investment).
. The future challenge will consist in finding the best combination between human and AI.

④ Substantive Fairness and AI

. The principle of fairness implies the commitment to ensure:
  . equal and just distribution of benefits and costs;
  . that individuals and groups are free from unfair bias, discrimination and stigmatisation.
. Depending on the specific domain of application, the concept of fairness can have a slightly different meaning.
. In the AI decision making domain, the substantive dimension (as also specified in GDPR) concerns the so-called informational fairness (strictly connected to the principle of transparency since it requires individual to be informed about the automated processing) and also the fairness of the content of the inference (e.g. inference should avoid prejudice, discrimination etc.), with regard to:
  . appropriate mathematical/statistical procedures for profiling;
  . technical and organisational measures to ensure correctness of personal data;
  . secure personal data (by taking into account potential risks and discriminatory effects).

⑤ The COMPAS system

- COMPAS is an actual risk assessment tool used by US judges to determine the risk of recidivism and consequently suggest an appropriate correction treatment.
- It is based on statistical algorithms to establish risk profiles associated with groups of individuals sharing particular features.
- Offenders are classified, based on probability scores, in 3 categories (high, medium and low risk).
- The score is based on:
  - multiple-choice test (that most of the defendants are required to do after being arrested the first time);
  - static risk variables (e.g. prior criminal history, education level, etc.);
  - dynamic risk variables (e.g. drug abuse, employment status, social integration, etc.).
- The Loomis case:
  - in 2013 E. Loomis was charged with driving a stolen vehicle and fleeing from police;
  - the Distrectual Court ordered a pre-sentencing investigation that included the COMPAS risk assessment;
  - Loomis was classified at high risk for recidivism and sentenced to 6 years imprisonment (a particularly severe sentence for his crime);
  - the decision was appealed by Loomis for violation of due process rights of defence, arguing that COMPAS is a proprietary software and thus its functioning is unknown and its validity cannot be verified;
  - moreover, he argued that it discriminates on gender and race, and that statistical-based predictions violate the right to individualized decision;
  - in 2016 the Supreme Court of Wisconsin rejected all defendant's arguments, since according to them:
    - statistical algorithms operate on the basis of generalisation (i.e. by comparing some features of the defendant with the same features of similar individuals), but they does not violate the right to individualized decisions;
    - they should be used to enhance a judge's evaluation of other evidence in the formulation of an individualized sentencing;
    - the prohibition to base decisions solely on risk scores, together with the obligation to motivate the verdict, should safeguard the defendants' rights;
    - the Court also rejected the argument according to which COMPAS discriminates against men, since gender is a necessary feature to achieve statistical accuracy;
    - finally, the Court established that judges should be informed on the debate concerning COMPAS race discrimination.
- The debate:
  - since the Loomis case, the use of COMPAS has been highly debated with regard to both its fairness and accuracy;
  - in 2016 ProPublica published a study involving ~11000 defendants assessed by COMPAS to evaluate its accuracy and fairness by comparing the predicted and actual recidivism rates;
  - the results of the study revealed:
    - a moderate-low predictive accuracy (61.2%);
    - that black defendants were predicted at a higher risk than they actually were, with a probability of high-risk misclassification of 45% (vs. 23% for whites);
    - that, conversely, white defendants were predicted to be less risky than they were, with a probability of low-risk misclassification of 48% (vs. 28% for blacks).
  - a second study by Northpoint argued that ProPublica made several statistical and technical errors, such as a mis-specified regression model, and wrongly defined classification metrics;
  - in particular, it showed that:
    - the accuracy of COMPAS predictions is higher than the one of human judgements;
    - COMPAS is compliant with fairness principles and it does not implement racial discrimination;
    - in fact, the higher recidivism predictions for blacks is due to the different base rate within the group;

· in particular, the prediction is equally accurate for both groups.

· To actually assess the fairness of COMPAS, we can develop and analyse a toy problem (which we call SAPMOC):

· 2000 defendants (1000 blues and 1000 greens);

· a single predictor (previous offences ⇒ probable recidivism);

· assumption 1
  ↳ previous offenders ⇒ 75% recidivate
  ↳ first-time offenders ⇒ 25% recidivate

· assumption 2
  ↳ blue ⇒ 75% previous offenders
  ↳ green ⇒ 25% previous offenders

	Recidivism	No recidivism	Total
Previous offence	750	250	1000
No previous offence	250	750	1000

Real outcomes

	Recidivism	No recidivism	Total
Previous offence	1000	0	1000
No previous offence	0	1000	1000

SAPMOC predictions

Blue:
TP = 750 · 75% = 562.5
FP = 750 · 25% = 187.5
TN = 250 · 75% = 187.5
FN = 250 · 25% = 62.5

Green:
TP = 250 · 75% = 187.5
FP = 250 · 25% = 62.5
TN = 750 · 75% = 562.5
FN = 750 · 25% = 187.5

Base rates	Positives (TP+FN)/Total	Negatives (TN+FP)/Total
Blue	62.5%	37.5%
Green	37.5%	62.5%

	SAPMOC accuracy (TP+TN)/Total
Blue	75%
Green	75%

· as we can see, the accuracy is equal for both groups;

· concerning the fairness, we can evaluate 5 criteria:

  · statistical parity ⇒ each group should have an equal proportion of positive and negative predictions

Statistical parity	Positives (TP+FP)/Total	Negatives (TN+FN)/Total
Blue	75%	25%
Green	25%	75%

⇒ ⊗

  · equality of opportunity ⇒ members of each group sharing the same features should be treated equally in equal proportion

Equality of opportunity	Positives TP/(TP+FN)	Negatives TN/(TN+FP)
Blue	90%	50%
Green	50%	90%

⇒ ⊗

  · calibration ⇒ the proportion of correct predictions should be equal within each group and with regard to each class

calibration	Positives TP/(TP+FP)	Negatives TN/(TN+FN)	→ ✓
Blue	75%	75%	
Green	75%	75%	

- Conditional use error ⇒ the proportion between FP (resp. FN) and the total amout of positive (resp. negative) predictions should be equal for the 2 groups

False rate	Positives FP/(FP+TP)	Negatives FN/(FN+TN)	⇒ ✓
Blue	25%	25%	
Green	25%	25%	

- Treatment equality ⇒ the ratio between errors in positive and negative predictions should be equal in all groups

Treatment equality	Positives FP/FN	Negatives FN/FP	⇒ ⊗
Blue	300%	33,3%	
Green	33,3%	300%	

- Considerations on SAPMOS:
  - equal accuracy within groups;
  - different base rate explains the violation of statistical parity, treatment equality, and equality of opportunities;
  - violation of fairness criteria does not necessarily lead to actual unfairness;
  - imposing statistical parity would result in lower accuracy, higher false rate and discrimination against individuals;
  - thus, there is a trade-off between individual fairness and group fairness.

⑥ Considerations

- We can think of a decision-making process as a complex process consisting of different phases.
- A prediction is not a decision, since making a decision requires applying judgement to a prediction and then acting upon such judgement.
- Unfairness may happen in every phase:
  - unfairness in prediction (prohibited features, biased data set, biased proxy, etc.);
  - unfairness in classification (different base rates, which can be addressed with the so-called affirmative actions - giving priority to the under-represented group - or with different thresholds);*
  - unfairness in decision (right and values optimization).

* Obs. affirmative actions are equivalent to using different thresholds, but the formers do not require to modify the prediction while the latter does.

- AI is too often perceived as a source of threats, while Law is too often seen as difficult and sometimes even inaccessible for citizens.
- The combination of AI and Law could be the key to protect citizens and make the Law accessible to the wider public.

→ 11. AUTONOMOUS VEHICLES:

Autonomous Driving - Ethical & Social issues (talk):

There is an intimate connection between autonomous driving and ethics.

① Autonomous Vehicles

- AVs look like normal vehicles with the addition of several sensors, both LIDAR and cameras.
- In particular, Waymo's AVs are capable of having a 360° vision, and are trained by driving the car in busy cities.
- Waymo's AVs are also capable of predicting others' movements, and of preventing them from getting hurt.
- The degree of automation is very important, and it's a ladder rather than an "all or nothing" situation ⇒ to develop an advanced AV we should start from a basic vehicle and add functions progressively.
- Such degrees of automation were introduced in the SAE J3016 document:
  - L0
  - L1
  - L2

  } the user is driving

  - L3
  - L4
  - L5

  } the user is not driving

  Full automation: every person in the car is an actual passenger → the vehicle can operate autonomously under certain conditions (if these aren't met, the control is given back to the user)

② Ethics of AVs

- Level 5 autonomy is a technical goal, but it brings out many social issues and questions:
  - Why do we want AVs?
    - ↳ autonomy and freedom
    - ↳ safety
    - ↳ sustainability
    - ↳ inclusiveness
  - Who would use AVs?
    - ↳ everybody
    - ↳ only those who can afford them
    - ↳ only those who can take control, if needed
  - How would we accomplish them?
    - ↳ social trust
    - ↳ responsibility
    - ↳ rights
    - ↳ by Law/discretionary
  - Where would AVs operate?
    - ↳ everywhere
    - ↳ highways
    - ↳ parking lots
    - ↳ national/international
- Thus, L5 is not only a technological goal but it will be negotiated taking into account society as a whole, ethics and regulations.
- Main ethical issues:
  1. Unavoidable collisions
  2. Privacy and security
  3. Responsibility allocation
  4. Sustainability
  5. Freedom and the Social Good
- Design is in the middle between our technical goal and all the social factors, and it's the way in which values are implemented.
- Initial recommendations have been written in EU ("Ethics of Connected and Automated Vehicles, September 2020), which serve as a baseline for future European regulations and provide

an ethical framework.

2.1 Unavoidable collisions

- Unavoidable collisions have been a primary worry in the ethical debate on AVs, since they're situations in which harm is inevitable but can be distributed in different ways.
- We can develop accident algorithms to program the AV software s.t. they can make a choice when they find themselves in such situations (such choices are ethical choices).
- The main issues are deciding which ethical theory implement, how that can be done, and who gets to decide.
- If such decisions are made only by the car manufacturer, then the personal autonomy of the driver is violated.
- On the other hand, if such decisions are made only by the driver, then the rights of bystanders may be neglected.
- Moreover, some people argue that the debate on unavoidable collisions is a bit detached from vehycle dynamics, available/foreseeable tech and realistic scenarios.
- Still, this remains an open problem and deserves to be analysed.

2.2 Privacy and Security

- For AVs to function properly, a huge quantity of data must be collected, shared and stored.
- We need new definitions of privacy and security of sensible data as involved in AD.
- We also need to understand which data must be protected and garantee data protection throughout the whole infrastracture (and not only at the car level).
- Consent is also problematic, since the AV's sensors are collecting data also related to other vehicles and bystanders.
- Therefore, in addition to standard mechanical vehicle safety we need to guarantee also a digital safety against external attacks, software issues and data thefts/leaks.
- Cybersecurity is particularly important to contrast attacks (e.g. stickers on road signs).

2.3 Responsibility allocation

- This issue is related to who should be held responsible for harm caused by accident involving AVs.
- According to the Meaningful Human Control approach, AVs must be designed and deployed in a way that assures a satisfying exercise of human moral responsibility, and a clear and fair distribution of legal liability.
- Such approach, however, has a huge impact on L5 automation.

2.4 Sustainability

- Environmental impact:
  - there's the common belief that with AVs we will be able to better handle traffic and thus reduce pollution;
  - however, someone argues that, by contrast, there will be more cars in use and thus it will be more difficult to handle traffic;
  - other factors, as the reusability and recyclability of materials and energy consumption (especially for data centres), should be considered.
- Social impact:
  - it is not clear whether there will be more or less traffic;
  - a good aspect is that probably people with disabilities will be included.
- Economic impact:
  - as with many technological innovations, there is the risk of job losses, and maybe the need of new jobs will arise;
  - someone also agued that it would be better to lower the price of AVs to make them more accessible to middle and lower classes (for fairness)

8.5 Personal freedom and the Social Good

- On one hand, having many AVs would lead to more safety in roads, but on the other hand it would undermine the pleasure of driving.
- There are also other value conflicts, as personal privacy vs system efficiency, and moral autonomy vs human error.
- Outlawing human driving would:
  - minimise road casualties ⇒ ✓
  - maximise traffic efficiency ⇒ ✓
  - undermine individual freedom ⇒ ✗
  - introduce discrimination ⇒ ✗

→ 12. THE CLAUDETTE SYSTEM:

The Claudette system - Automation of personal data and consumer law enforcement using AI (talk):

"Claudette" stands for "clause detector": it is a research project which aims at automating the possibility of enforcing personal data and consumers law using AI.

① AI and Law

- Nowadays, we reveal much more information about ourselves than ever before, and such information are being collected by companies which analyse them using AI.
- There are many concerns related to data privacy, and as a result AI is seen as something at the service of businesses.
- However, AI can unlock consumer-empowering technologies, such as:
  - protection against unwanted monitoring (GDPR);
  - support in detecting unfair use of AI;
  - control commercial practice fairness.
- Claudette is a ML system, based on supervised learning, which automatically detects potentially unfair clauses in Terms of Services and Privacy Policies:
  - consumers tend to agree without actually reading;
  - NGOs have competence to control but lack resources;
  - businesses keep using unlawful clauses.

② The dataset

- The system was trained on an initial dataset of 50 manually annotated Terms of Services (ToS), with 11,1% positives (i.e. potentially unfair clauses).
- The current dataset consists of 100 ToS.
- The data is labelled based on Directive 93/13, Article 3.1: "A contractual term which has not been individually negotiated shall be regarded as unfair if, contrary to the requirement of good faith, it causes a significant imbalance in the parties' rights and obligations arising under the contract, to the detriment of the consumer.
- In practice, there are some type of clauses that traders are prohibited from using in the contracts (e.g. arbitration, unilateral change, content removal, limitation of liability, etc.), classified in 3 tiers of unfairness and identified by a specific XML tag.

③ Unfair clauses

- Consent: if a clause states that the consumer is bound by the ToS simply by visiting the website or by downloading the app, it is potentially unfair (e.g. Airbnb, Facebook)
- Jurisdiction:
  - if giving consumers a right to bring disputes in their place of residence, it is clearly fair;
  - if stating that any judicial proceedings takes a residence away (i.e. in a different city or country), it is clearly unfair (e.g. Dropbox).
- Limitation of liability:
  - if stating that the provider may be liable, it is clearly fair (e.g. World of Warcraft);

- if stating that the provider will never be liable for any action taken by other people// damages incurred by the computer because of malware// when contains a blanket phrase like "to the fullest extent permissible by law", it is potentially unfair (e.g. 9gag);
- if stating that the provider will never be liable for physical injuries (health/life)// gross negligence// intentional damage, it is clearly unfair (e.g. Rovio).

④ ML methodology

- The problem was modelled as:
  - a detection task ⇒ check whether a sentence contains a potentially unfair clause (positive if unfair, negative otherwise);
  - a sentence classification task ⇒ determing the category the unfair clause belongs to.
- Appraches:
  - Bag of Words (BoW), to leverage lexical information in sentences;
  - Tree kernels, to describe grammatical relations between words in a sentence through a tree;
  - CNNs, SVM, etc.
- The algorithm was trained via the Leave-One-Out procedure, namely each document in turn is used as test set, leaving the remaining documents for training set (4/5) and validation set (1/5) for model selection.
- Three metrics were employed:
  - Precision ⇒ fraction of positive predictions actually labelled as positive;
  - Recall ⇒ fraction of positive examples that are correctly detected;
  - F1 ⇒ harmonic mean between precision and recall.
- A random classifier was used as a baseline for comparison.
- After training various models on 50 ToS, it was observed that the best performing system is an ensemble.
- Claudette correctly detected around 80% of the potentially unfair clauses in each category, from a minimum of 72.7% for arbitration clauses to a maximum of 89.7% for jurisdiction clauses (however the dataset was highly imbalanced, and the lower performance for arbitration clauses may be justified by the fact that there are fewer of them).

⑤ Claudette online server

- The system is accessible via an online server.
- Given an input ToS, it analyses it and shows all the detected potentially unfair clauses together with the predicted category and confidence score.
- It can also provide a rationale justifying the decision:
  - Human legal experts are able to recognize potentially unfair clauses thanks to their background knowledge of the domain:
    - they rely on intuitions, trained on experience with relevant examples;
    - they are able to explain their intuitions of unfairness, provide reasons why a clause is unfair (Legal Rationales), and use rationales to guide such intuitions;
    - they appeal to their background knowledge as support for reasoning.
  - To develop such functionality in Claudette, Memory-Augmented Neural Networks (MANN) were employed:
    - they process input data and store the information in some memory;
    - they understand pieces of knowledge relevant to a given query;
    - they retrieve concepts from memory;
    - they combine memory and query to make a prediction.
- Experimental setup for unfairness identification:
  - First of all the knowledge base was developed as a list of rationales from human legal experts.
  - Given an input clause from a ToS to be classified, the system retrieves from the KB the most similar rationale (based on a similarity score).
  - The input clause is aggregated with the rationale, forming a kind of enhanced input.
  - The enhanced input is used once again as an input for another iterations to refine it.
  - Once all the relevant content has been extracted, the enhanced input is fed to an Answer

Module which makes a prediction (Unfair/Other).

⑥ Claudette meets GDPR

- A perfectly compliant privacy policy should reflect the so-called Golden Standard:
  - Lawfulness.
  - Fairness.
  - Transparency.
- Three dimensions of compliancy:
  1. Comprehensiveness of information ⇒ the policy should contain all the information required by articles 13 and 14 of GDPR (e.g. the name of the Controller, the purposes for which data are collected, etc.).
  2. Substantive compliance ⇒ the policy should only allow for processings of personal data that are compliant with GDPR.
  3. Clarity of expression ⇒ the policy should be framed in an understandable and precise language (e.g. it should not be possible to have different, conflicting interpretation of a clause).
- Different levels of achievement (optimal/suboptimal).

6.1 Comprehensiveness of information

- 23 categories for clauses, identified by XML tags and classified with a numeric value (as in ToS classification).
- Categories of personal data concerned:
  - if such categories are comprehensively specified and not vague, it is fully informative (e.g. Google Privacy Policy);
  - in other cases (e.g. when a clause only provides examples), it is insufficiently informative.

6.2 Substantive compliance

- 10 categories for clauses, identified by XML tags and classified with a numeric value (as in ToS and Comprehensiveness of information classification).
- Policy change:
  - when notice is given and new consent is required, it is a fair processing;
  - when notice is given but a new consent (or confirmation of reading) is not required, it is a problematic processing (e.g. Twitter Privacy Policy);
  - when no notice is given and new consent is not required, it is an unfair processing (e.g. Booking Privacy Policy).

6.3 Clarity of expression

- 4 main indicators of vagueness:
  1. Conditional Terms => performance of a stated action or activity is dependent on a variable trigger, recognizable via qualifiers such as "depending", "as necessary", "as appropriate", etc.
  2. Generalization ⇒ terms vaguely abstract information practices using contexts that are unclear, recognizable via qualifiers such as "generally", "mostly", "commonly", etc.
  3. Modality ⇒ clause includes modal verbs, adverbs and non-specific adjectives which create uncertainty w.r.t. actual actions, recognizable via qualifiers such as "may", "might", "could", etc.
  4. Non-specific numeric quantifiers → they create ambiguity as to the actual measure, recognizable via qualifiers such as "certain", "numerous", "some", etc.
- There may be also combinations of different forms of vagueness

Obs. the performances of Claudette on GDPR are still not as good as those of Claudette on ToS, thus many algorithms are currently being tested.

Obs. also a web-crawler has been developed to check the date of a privacy policy and compare the content with a previous saved version, and thus to detect unnotified updates.

⑦ Future developments

- Experiment new methods for the assessment or privacy policies.
- Multilingualism [redacted] via transfer learning.
- Empowerment through transparency (linguistic transparency and explanations).

→ 13. INTELLIGENT WEAPONS:

Autonomous Weapons Systems:

Notions of autonomy:
- a capability that enables a particular action of a system to be automatic or, within programmed boundaries, self-governing (US Military Defense Science Board);
- the capacity to operate in the real-world environment without any form of external control, once the machine is activated for extended periods of time (George A. Berkley, a roboticist);
- an agent's capacity to learn what it can to compensate for partial or incorrect prior knowledge (Russel & Norvig);
- a system's capacity to perceive and interpret its environment, define and select what stimuli to take into consideration, according to its internal states (Castelfranchi & Falcone).

① The concept of autonomy

- If the standard is too high (i.e. to be considered autonomous, a system must replicate all the cognitive capacities of humans), then no artificial entity is autonomous.
- However, if it's too low then every algorithm can be considered autonomous.
- Autonomy can be seen as a scalable capacity, merging three dimensions:
  1. Independence;
  2. Cognitive skill;
  3. Teleonomic cognitive architecture (namely a cognitive architecture enabling the agent to pursue its goals).

1.1 Independence

- A technological device, within a system, is independent to the extent that it is able to accomplish a high level task on its own, without external interventions (e.g. land mine, collision-avoidance system in airplanes, etc.).
- There may be various levels of independence.
- In the case of aviation, systems are so complex that automation plays a large role in their management (still in combination with human intervention).
- Independence within a socio-technical system:
  - an integrated combination of human, technological and organizational components (e.g. to manage an airport you need to take into account airplanes, manned flying aircrafts as hybrids, and civil aviation);
  - there are also components of such systems that work autonomously (e.g. collision-avoidance systems, autopilot, etc.).

1.2 Cognitive skills

- An autonomous system engages in high-level cognition (involving the ability to discriminate facts, actions or outcomes) using its own abilities in one or more of the following ways:
  - acquisition and classification of input data;
  - information analysis to extract further information from input data;
  - action selection or construction of plans of actions from the extracted information;
  - action implementation.

E.g. A land mine is independent (i.e. it works on its own) but not autonomous (i.e. no cognitive skills), whereas drones such as Taranis are able to identify targets and avoid threats (it waits for a human confirmation of the target before shooting).

- In the case of a drone:
  - acquisition and classification of input data ⇒ acquiring input data from sensors, applying noise reduction and filtering;
  - information analysis ⇒ computing expected flight trajectories or possible encounters and alerting the operator of possible risks (e.g. bad weather or approaching objects);
  - decision and action selection ⇒ providing suggestions or list of options, or taking action;
  - plan implementation and monitoring ⇒ flying according to the established route and monitoring missiles.
- Humans in the loop:
  - autonomy of a device increases as the device is delegated a larger share of the required cognitive tasks;
    - ↳ increased independence of the device;
    - ↳ increased interaction/collaboration between the human and the artificial component;
  - humans may remain in the loop while technological devices execute the larger share of the cognitive functions involved in the performance of the task;
  - the delegator chooses to delegate choices instrumental to the execution of a function to the cognitive skills of the delegatee system (e.g. flying aircraft, engaging target);
  - the delegator does not know, and thus does not intentionally pre-select, what the delegated system will choose to do in future situations (e.g. how to fly, what particular target to engage).

1.3 Cognitive-behavioural architecture

- Concept related to:
  - adaptiveness (auto-teleonomy);
  - teleology (purposiveness) and intentionality.
- An adaptive agent can change its patterns of behaviour to better achieve its purposes, in the environment in which it operates:
  - it interacts with the environment, getting inputs and providing outputs;
  - on the basis of environmental inputs, it changes the internal states on which its behaviour depends.
- It has a feedback or homeostatic mechanism, which keeps the sistem focused on its objective by changing its internal state as the environment changes, and so enabling the system to act as required by the changed environment.
- A teleologic system has explicit cognitive states:
  - goals;
  - beliefs;
  - plans;
  - intentions.
- Such cognitive states are implemented differently from the more complex corresponding human mental states, but they perform the same basic functions.

E.g. A drone flies to the target zone, identifies the target, and select and implement a strategy to eliminate it: it has an internally stored representation of its goals, it acquires inputs from the environment, it develops and implements flight plans to reach the target.

② Other types of autonomy

- Collective adaptiveness:
  - drones flying in a flock, where the persistence of the flight formation is only determined by the fact that each drone keeps a certain distance from the others;
  - a set of land vehicles involved in the elimination of landmines may cover all of the area to be cleaned since each of them follow certain simple rules concerning movements and distances from the other.
- Multilayered autonomy:
  - the autonomous behaviour of a system may also emerge from the interaction of lower-level non-autonomous or autonomous elements (e.g. the adaptation of evolutionary algorithms resulting from the higher combination of "genes");
  - agents may be flexibly integrated into higher units of agency through information and de-

cision sharing.

③ Autonomy in weapon systems

- An agreement on regulating lethal AWS has not yet been found, but there is pressure to promote treaties banning or limiting the use of such systems.
- In 2012, US issued the Directive on Autonomy in Weapons Systems, focusing on target selection and on the distinction between autonomous and semi-autonomous weapons:
  - target selection involves "the determination that an individual target or a group of targets is to be engaged";
  - autonomous weapons are those that "once activated, can select and engage targets without further intervention by a human operator";
  - conversely, semi-autonomous weapons are "intended to only engage individual targets or specific target groups that have been selected by a human operator".
- autonomous weapons should only be used to apply non-lethal, non-kinetic force (they may engage, under human supervision, non-human targets for the defence of manned platforms);
- semi-autonomous weapons may be deployed for any purpose, including the exercise of lethal force against humans, subject only to certification.

Obs.: Israeli's Iron Dome is an autonomous weapon system to intercept and destroy incoming missiles.

- Criticisms to this distinction:
  - there are two phases in the targeting process involving semi-autonomous weapons:
    - first humans delimit the domain of the targets to be selected (go-onto-target), or a location is space (go-onto-location-in space);
    - then, it's up to the machine to select what particular object to engage within that domain/location.
  - thus, in a certain sense it's the machine that really decides the target to engage (even in semi-autonomous scenarios), as a result of a double choice (human operator giving a general description of the target + machine locking on a specific object).
- Weapons may also rely on a cognitive architecture (i.e. the teleological ability to develop plans on how to detect and engage the target, given the available information).

E.g. The long-range antiship missile by Lockheed-Martin can reroute around unexpected threats, search for an enemy fleet, identify the one ship it will attack among others in its vicinity, and plan its final approach to defeat anti-missile systems, all out of contact with any human decision maker (but possibly in contact with other missiles, which can work together as a team).

- The targeting process includes all aspects of decision making, which can be (partially or totally) automated.
- Different kinds of responsibility:
  - functional responsibility ⇒ what defect caused the harm;
  - blameworthiness ⇒ whether the failure that caused the harm involved a fault, namely a substandard behaviour in a moral agent;
  - legal liability ⇒ who is legally liable for tort.
- Two aspects of the laws of war:
  - Jus ad bellum ⇒ regulations expressing when it is justifiable to declare war against other country (e.g. the United Nations Treaty prohibits wars of aggression and allows only wars for defence, even though this distinction is challenged by humanitarian wars).
  - Jus in bellum ⇒ principles expressing how we should behave when in war.
    - ↳ Necessity ⇒ harm inflicted to enemies must be justified by a purpose;
    - ↳ Distinction ⇒ militar activity should be directed against enemy army and not civilians;
    - ↳ Proportionality ⇒ the harm inflicted to civilians (admitted as a side effect) should be proportional to the pursued military goal.

- Some argue that machines would not be subjected to human emotions, as regular soldiers are (e.g. revenge acts).
- On the other hand, a machine may not be able to distinguish a [redacted] soldier by a civilian.
- Moreover, autonomous weapons could not only trigger a new arms race but also facilitate richer countries, which could invest more money in the development of AWS.
- symbiotic man-machine partnership:
  - we should aim at a symbiotic partnership between human and machines, which would perform intellectual operations much more effectively than man or machine alone;
  - we could reject machine autonomy and retain machine cognition for exploratory, [redacted] constraining or implementation functions.
- Liability gap:
  - it's impossible to attribute moral responsibilities (blameworthiness) and legal liabilities to anyone for certain harms caused by the system's autonomous operation;
  - especially serious in the military domain (but also in the civil domain).

→ 14. Ethics of Filtering:

Ethics of filtering (talk):

The Digital Services Act (DSA) will replace the old eCommerce regulations, and will try to regulate all those online platforms and digital services in which users exchange information. Users commonly interact with each other via user-generated content, which:
- enables users to express themselves;
- lets users create, transmit or access information and cultural creations;
- lets users to engage in social interactions.

① Moderation and filtering

- since there are many interactions between users, moderation becomes mandatory.
- Moderation is the active governance of platforms meant to ensure interactions among the users that are:
  - productive;
  - pro-social;
  - lawful.
- Sometime filtering is needed to:
  - prevent unlawful and harmful online behaviour;
  - mitigate its effects;
  - facilitate cooperation;
  - prevent abuse.
- We can identify a taxonomy for the different filtering methodologies:
  1. Where
     - centralized
     - distributed
  2. When
     - ex-ante
     - ex-post
       - reactive
       - proactive
  3. How
     - transparently
       - contestable
       - non-contestable
     - secretly
  4. Who
     - manually
     - automatically
     - hybrid

Obs. finding the right trade-off between moderation and censorship is not an easy task.

1.1 Taxonomy - Where

- Centralized filtering ⇒ it is applied by a central authority according to uniform policies, that apply to a whole platform.
- Decentralized filtering ⇒ it involves multiple distributed moderators operating with a degree of independence, and possibly enforcing different policies on subsets of the platform.

1.2 Taxonomy - When

- Ex-ante filtering ⇒ it is applied before the content has been made available on the platform.
- Ex-post filtering ⇒ it is applied on the content that is already accessible to the platform's users
  - ↳ reactive, if the filter takes place after the issue has been signaled by users or third parties;
  - ↳ proactive, if the filter takes place upon initiative of the moderation system, which thus has the task of identifying the issue.

1.3 Taxonomy - How

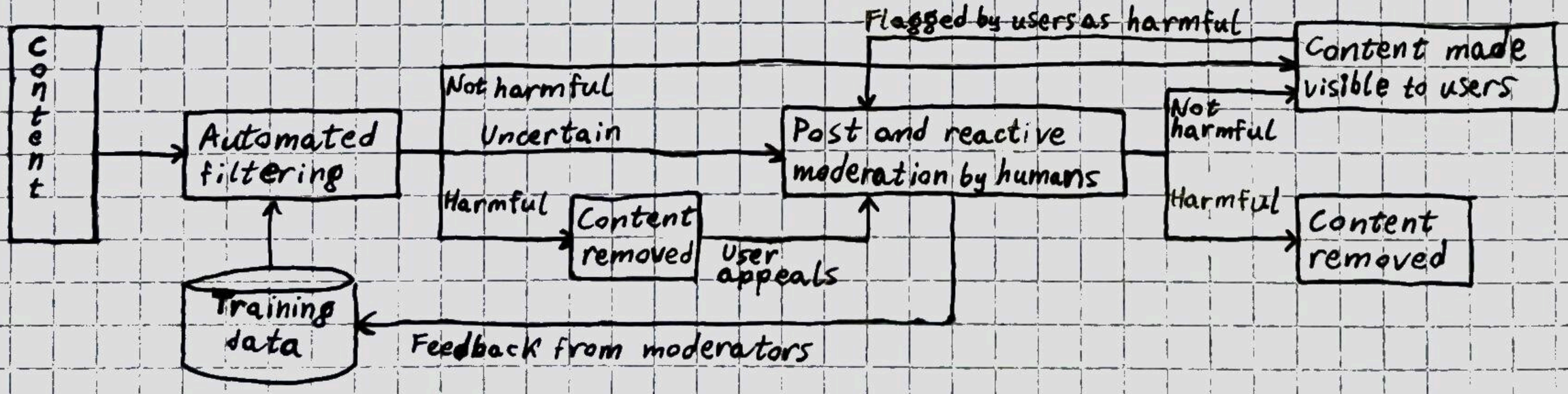
- Transparent filtering ⇒ it provides information on the exclusion of items from the platform.
  - ↳ contestable, if the platform provides uploaders with ways to contest the outcome of the filtering, and to obtain a new decision on the matter;
  - ↳ non-contestable, if there is no remedy available to the uploaders.
- Secret filtering ⇒ it does not provide any information about the operation.

1.4 Taxonomy - Who

- Manual filtering ⇒ it is performed by humans.
- Automated filtering ⇒ it is performed by algorithmic tools.
- Hybrid filtering ⇒ it is performed by a combination of humans and automated tools.

② High-level working

- Filtering can work on different media, from text to audio to images and videos.
- The filtering algorithm changes according to the kind of media to be filtered:
  - metadata searching, hashing and fingerprinting ⇒ to identify copies of known digital works;
  - blacklisting ⇒ to find unwanted expressions;
  - NLP ⇒ to address meaning and context;
  - multiple AI techniques ⇒ to identify unwanted images, or combinations of text and images, and to translate spoken language into text.



③ Limits of automated filtering

- Filtering algorithms (based on ML) most of the time lack commonsense:
  - Facebook removed an image of Copenhagen's Little Mermaid because of nudity;
  - for the same reason, it also removed images of Bologna's Neptune.
- In some cases, they may also apply double standards:
  - when the New Zeland terrorist attack was broadcasted on the internet and went viral, the main social networks initially didn't banned the video because their filtering systems weren't able to spot its content;

- conversely, YouTube frequently removes videos about war in Syria, which should instead be considered "safe".

④ Santa Clara Principles

- One of the initiatives aiming at pushing companies into being more transparent about the filtering procedure.
- Three principles:
  - Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.
  - Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.
  - companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.
- Companies started releasing a transparency report, which shows statistics for the suspended accounts and removed contents.

⑤ Other issues

- Filter bubbles.
- Echo chambers.
- Censorship.
- Fake news.

→15. A FRAMEWORK FOR ETHICAL PRINCIPLES:

AI Ethics at IBM - From Principles to Practice (talk):

There are many principles around AI ethics that have been put together, and such principles have been put in practice at IBM.
Many concerns that people have about AI are actually specific to some AI techniques and not all of them.
The popularity of ML and DL, nowadays, is related to the abundance of data and of (in many different fields) and of computing power. This leads to a huge success in CV and NLP, in which the capabilities of machines have outperformed the capabilities of humans
AI has however many limitations:
- Narrow AI (i.e. it can solve only very specific problems).
- Lack of robustness and adaptability (e.g. adversarial attacks).
- It needs lots of resources (data and computing power).

① Ethical issues

- Examples of ethical issues related to AI:
  - gender-biased Apple credit card approval process;
  - discrimination in Uber and Lyft ride-sharing dynamic pricing;
  - gender-biased Amazon recruitment software;
  - Microsoft chatbot exhibiting racist speech;
  - Facebook and Cambridge Analytica's unhetical usage of personal data.
- Some may argue whether we can trust AI's decisions.
- AI ethics:
  - multidisciplinary field of study (computer science + economy + phylosophy + law);
  - it focuses on optimizing AI's beneficial impact while reducing risks and adverse outcomes;
  - it also aims at designing and building AI systems that are aware of the values and principles to be followed in the deployment scenarios;
  - finally, it focuses on identifying, study and propose also non-technical solutions for ethics issues arising from the pervasive use of AI in life and society.

- Main ethics issues:
  - AI needs data ⇒ data privacy and governance.
  - AI is often a black box ⇒ explainability and transparency.
  - AI can make/recommend decisions ⇒ fairness and value alignment.
  - AI is based on statistics and has always a small percentage of error ⇒ accountability in case of mistakes.
  - AI can profile people and manipulate their preferences ⇒ human and moral agency.
  - AI is very pervasive and dynamic ⇒ larger negative impacts for tech misuse, and fast transformation of jobs and society.
  - Good or bad use of technology ⇒ autonomous weapons and mass surveillance, and UN Sustainable Development Goals.
- AI is not a neutral technology:
  - misuse must be avoided;
  - it needs to be designed and developed with the right properties (i.e. fair, explainable, etc.).

② Fairness and bias

- AI may be biased, and have a prejudice embedded into itself (e.g. biased dataset).
- Thus, one could behave unfairly to certain groups compared to others.

E.g. Imagenet, consisting in 14M images, presents a bias in data distribution and labels (due to MturK people).

- Bias is not just in the training data, but also is design decisions.

E.g. Mortgage application: apart from a correlation gender-acceptance in training data, the prioritized motivations for loan applications introduced bias (e.g. buying a house, paying school fees, paying legal fees).

- The definition of fairness is somehow challenged by recidivism assessment systems:
  - overall accuracy is the same regardless of race;
  - likelihood of recidivism among defendants labelled as medium or high risk is similar, regardless of race;
  - but still, false positive and false negatives rates are very different.
- Many decision points:
  - individual vs group fairness ⇒ similar individuals should receive similar treatment vs. groups defined by protected attributes should receive similar treatments;
  - more context-dependent definitions of fairness;
  - acceptable bias threashold (typically 80%, but varies according to the specific domain);
  - when to detect bias (in the training data or in the learnt model);

Obs. all these decisions are made by humans, who must be educated on what bias is and how to avoid it.

③ Explainability

- AI systems cannot be black boxes, they must be able to provide explanations for their decisions.
- The GDPR in fact requires that the data subject has the right to be provided with a meaningful information about the logic involved in a decision based on automated processing.
- The system should also be able to provide different explanations for the same decision, depending on the target audience.

④ Profiling and manipulations

- Not only our preferences are inferred from our actions online (e.g. likes, images, follows, etc.), but also they are tried to be made simpler and more polarized.
- In fact, simple preferences are better for targeted ads.

⑤ Impact on the workforce

. Many jobs will disappear, but many others will be created.
. All jobs will change:
  . some tasks may be delegated to a machine;
  . other tasks may still be performed by a human operator.

⑥ A vision of the future

. The UN has put together this vision for 2030 of what it means to improve the world.
. It includes 17 sustainable goals (e.g. no poverty, zero hunger, etc.).
. A study has shown how much current AI has been used to move towards these goals.
. COVID pandemic has worsened this situation.

⑦ IBM and AI ethics

· The business model of IBM is to deliver technology to other public and private companies.
· They developed many AI research projects, from Deep Blue to IBM Watson to Project Debater.
· IBM Principles of Trust and Transparency:
  · the purpose of AI is to augment human intelligence ⇒ solutions to support human decision-makers;
  · data and insights belong to their creator ⇒ data is not reused for other purposes;
  · new technology, including AI systems, must be transparent and explainable ⇒ the user should know whether he/she is interacting with a human or an AI.
· What it means to trust a machine-made decision:
  · it must be fair, s.t. it does not discriminate anyone;
  · it must be able to explain why it made a decision (not a black box);
  · it must be robust;
  · it must be transparent.
· AI fairness at IBM:
  · technical solutions to detect and mitigate AI bias:
    · research work;
    · open source libraries (e.g. AI fairness 360);
    · proprietary tools (e.g. Watson OpenScale);
  · developers' education and training:
    · AI bias education module for IBMers;
    · developers' awareness material;
    · revised methodologies for the AI pipeline;
    · adoption strategies;
    · governance frameworks;
    · consultations with stakeholders;
    · design thinking actions.
· AI transparency at IBM:
  · AI factsheet:
    · transparency by documentation;
    · design and development choices;
  · useful to:
    · developers;
    · clients;
    · users regulators/auditors;
  · aligned with EC High Level Expert Group on AI self-assessment list (ALTAI);
  · open source library (AI factsheet 360).
· From principles to guidelines
  · AI ethics principles and issues are put together with guidelines, toolkits and education under the governance of IBM to ensure that these principles are put in practice.
  · There are also external partnerships.
  · The governance is performed by the IBM AI ethics board:
    · mission:
      · awareness and coordination;

- internal education and retraining;
- linking research to services and platforms;
- advice to business units;
- internal governance framework;
- define policies and advice regulators;
- decide whether a solution can be given to a client or not.

- risk based approach for the BUs => vetting based on 3 dimensions (tech, use, client).

⑧ <u>Not just AI</u>

- Neurotechnologies:
  - huge potential for healthcare;
  - reading/writing neurodata;
  - additional issues around privacy, agency and identity.
- Quantum computing: responsible usage of such a huge computing power.