Ethics

Francesco Farinola

April 2021

Contents

1 Talk ICHMS

1.1 AI revolution: empowering whom?

The AI community is responsible for the introduction of Agents as '**autonomous**' and '**social**' cooperating with human by following true **norms** and critically adopting our goals (not just **executing**)

This obliges us to become aware of possible appropriation of their creations, of possible unacceptable uses of these instruments. Are we ready for the anthropological revolution which is *also an economic, social and political revolution.*

For the mass media, the **main problems** are privacy, security, fake news, misinformation, hackers' attacks, anthropomorphism, war and artificial soldiers/arms, ethics inside Artificial creatures and algorithms. Not less serious problems is the future of work in 4.0 economy.

1.2 Is research only business oriented?

Better for whom? It is *not a technical problem* but a *political* one. Better for poor and powerless people or for dominating classes, lobbies, power countries. We want a **beneficial** AI but we should choose on which side to be.

It can be very beneficial for democracy, good market with reduced deception and manipulation, for social planning and decision, or transparency and control, participation.

In 2016, Carnegie Mellon University plans to create a research center that focuses on the ethics of artificial intelligence, called *K&L Gates Endowment for Ethics and Computational Technologies.* They stated that an array of academic, governmental and private efforts are led to explore a technology that until recently was largely the stuff of science fiction. Again, here we see an alliance oriented only to business just because it needs money.

1.3 Hidden Interests and Awareness Technology

Also the goals of our agents and robots, are the explicit, transparent at least for us? They should be **trustworthy and comprehensible** and must be able to **explain us why** they do what they do and the **reasons** and **motives** of their actions, decisions, suggestions. This requires a cognitive model of reasons and motive for believing and for goal processing/decisions.

Interests theory: what is better for me and my goals but ... i don't understand or intentionally pursue them.

Tutelary Role theory: X takes care of my interests, of my good, even in conflict with me, with my current goals; X helps me or pushes me or obliges me!

Tutelary doesn't means protecting me, only caring of our individual personal interests, but also take care of: common interests, conflicts of interests, the 'commons' of public goods and their relevance and respect.

It my happen that recommender systems give us recommendations and suggestions by following a market criteria with a **personalized advertising**. They will decide for us but, instead of us or for our good?

1.4 Presences in our Mixed Reality and Society

Mouth of truth algorithm: clearly we are developing algorithms for ascertaining the truth. An algorithm for deciding about reliable sources, credible information, what is true among so many different claims and data. There will be dogmatic truths and undisputable authorities, like any culture? Which culture and values will be assumed as the right ones?

The autonomous and proactive intelligent entities will become 'presences' and 'roles' in our hybrid society and mixed and augmented reality. Which roles will those entities play in our life? Will it be our guardian angel with a tutelary role by helping, protecting, empowering us, or our Jiminy Cricket with its recommendations, our supervisor or our tempting spirit/devil for the benefit of some marketing policy or monopoly.

Will we listen to that moral and rational voice as our SuperEgo or will our SuperEgo be externalized (the voice of our mother/teacher). Both solutions will be there:

- **Social one**: externalized voices and agents (our best friend).
- **Reflexively social one**: an augmented internalized Self and Consciousness.

1.5 Disagreement Technologies

Conflicts: the presupposition of democracy, are not just conflicts of views and opinions, or due to different conceptions, information, reasoning. There are conflicts of **objective interests**. The problem is conflicts between interests of group and classes, or conflicts between private interests vs common interests, the 'commons' and public goods. Social conflicts don't have a verbal/technical solution but a political one. It is a matter of power and of prevailing interests and compromises.

Conflicts with their disagreements and agreements are thus the motor and principle of Democracy and of its possible effectiveness in changing society in favor of the submitted subjects, disadvantaged classes and group...viva conflicts!

Mark Twain stated that if voting made any difference they wouldn't let us do it, but the problem is much harder, it is not just a complot, is that we vote in a self-defeating way and our collective stupidity. Might political education and education to commons and digital society and participatory democracy be enough? They will help, but not in every minority (poor classes, ethnic groups).

Making conflicts to emerge and become aware of, making express disagreement, making transparent which interests are hidden and prevailing, should be (in democracy) one of the main tasks of intelligent social technologies.

Critical thinking: using web technologies for organizing movements is ok, but not so good *without promoting critical consciousness.* Not only by counteracting our confirmation bias, counteracting our tendency to gregariousness and bubble effect on the web, but by helping us to understand hidden powers and also our prejudices.

We need environments and agents for learning and developing a critical thinking attitude; to manage our cognitive and motivational biases. New intelligent and interacting technology shouldn't be used for selling and for dominating.

Demystifying the ideology of the NET: NET interaction is perceives as non hierarchical, without superstructure and mediation, individually managed, spontaneous, thus free.

Anti-manipulation: i would like to have not so much a personal virtual or robotic psychotherapist. Much more a life navigator in my main social role, but not a navigator telling me what to do. A tutor, inducing me to understand and to reflect about why I'm oriented in that direction. Making me conscious of who and how is persuading or just unconsciously manipulating me.

1.6 Conclusion

The great revolution of ICT and Big Data can give to society a glass were to observe themselves and follow what it is happening. A glass reflecting also what is invisible: hidden presences, the future (**A glass of the invisible**). To show hidden phenomena and interests.

Can we make visible the invisible hand and partially govern it? Can we overcome our alienation? Skeptical about that, worry about possible net-demagogy.

To see what is invisible: Artificially augmented awareness. Artifical Intelligence may either exploit or overcome our natural stupidity.

2 Ethics for Trustworthy AI

Trustworthy AI has three components which should be met throughout the system's entire life cycle:

- **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 components are necessary but *not sufficient* in itself to achieve Trustworthy AI. There are laws that permit killing (resisting aggression). Some unlawful examples: use AI systems to attack PC/website, deceive consumers. Unethical ex: manipulate people for changing their political idea. Non-robust uses: Autonomous weapons, cars.

AI systems need to be **human-centric**, resting on a commitment to their use in the service of humanity and the common good, with the goal of improving human welfare and freedom. Maximize the benefits of AI systems while at the same time preventing a minimizing their risks.

Ethics vs law: **Ethics** are norms indicating what should be done, with regard to all interest at stake:

- **Positive ethics**: norms shared in a society (social hierarchy, gender roles) - what people believe is good.
- **Critical ethics**: norms that are viewed as most appropriate or ration.

Law: norms that adopted through institutional processes and coercively enforced. There's a difference of speed between law and technology (law is slow and we must develop and fix technology)

Stakeholders committed towards achieving Trustworthy AI can voluntarily opt to use these Guidelines as a method to operationalize 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.

2.1 Lawful AI

It should comply with:

- EU primary law (the Treaties of the European Union and its Charter of Fundamental Rights)
- EU secondary law (regulations and directives, such as the General Data Protection Regulation, the Product Liability Directive, the Regulation on the Free Flow of 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)
- Laws of EU Member State laws (Italian law)

Laws can be *horizontal of domain-specific* rules (medical devices, music recommendation systems etc.).

AI ethics is a sub-field of applied ethics. 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.

2.2 Fundamental rights as a basis of trustworthy AI

- **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 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 exclusion. In an AI context, equality entails that the system's operations cannot generate unfairly biased outputs. (GS: we need to understand what this means)
- **Citizens' rights**: right to vote, the right to good administration or access to public documents, and the right to petition the administration.

2.3 Ethical principles

- **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.

- **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.

- **Fairness**:

  Substantive dimension:

  - Ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.
  - Promoting equal opportunity in terms of access to education, goods, services and technology.
  - 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.

- **Explicability**:

  To ensure *contestability*:

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

  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.

Methods of accountable deliberation to deal with such tensions should be established. Conflicts between prevention of harm and human autonomy. Also between welfare and security,, explicability and performance.

2.4 Requirements of Trustworthy AI

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* helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects (human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach + public controls).

2. **Technical robustness and safety**: AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm. *Resilience to attack and security*, protected against vulnerabilities that can allow them to be exploited by adversaries. *Fallback plan and general safety* (have safeguards that enable a fallback plan in case of problems). *Accuracy* (have the ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models). *Reliability and Reproducibility*: the results of AI systems should be reproducible, as well as reliable.

3. **Privacy and Data governance**: Prevention of harm necessitates 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 constructed biases, inaccuracies, errors and mistakes, malicious data. *Access to data*: data protocols governing data access should be put in place.

4. **Transparency**: closely linked with explicability. The datasets and the processes that yield the AI system's decision, should be documented (*traceability*). Technical processes of an AI system and related human decisions should be explainable (*explainability*). Humans have the right to be informed that they are interacting with an AI system (*communication*)

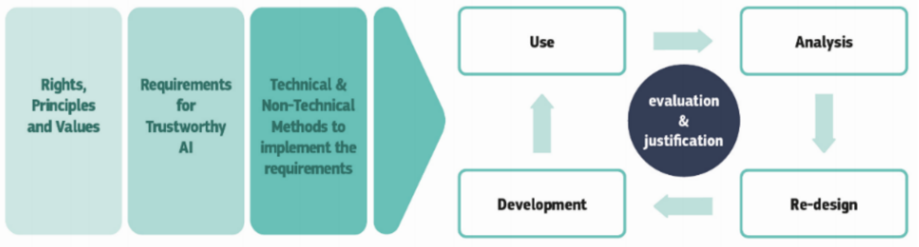
5. **Diversity, non-discrimination and fairness**: We must enable inclusion and diversity. *Avoidance of unfair bias*: Prevent unintended (in)direct prejudice and discrimination due to data or algorithms. *Accessibility and universal design*: AI systems should be user-centric and designed in a way that allows all people to use AI products or services, regardless of their age, gender, abilities or characteristics. *Stakeholder Participation*: Open discussion and the involvement of social partners and stakeholders, including the general public. *Diversity and inclusive design teams*: the teams

that design, develop, test and maintain, deploy and procure these systems reflect the diversity of users and of society in general.

6. **Societal and environmental well-being:** the broader society, other sentient being and the environment should be also considered as stakeholders throughout the AI system's life cycle. *Sustainable and environmentally friendly AI* should be encouraged. The effect of there systems on individuals and society must be carefully monitored (*social impact*). Take into account AI's effect on institutions, democracy at large.

7. **Accountability**: ensure responsibility and accountability for AI systems and their outcomes. Enablement of the assessment of algorithms, data and design processes (*auditability*). *Minimisation and reporting of negative impacts*. *Trade-offs* should be addresses in a ration and methodological manner within the state of the art. Accessible mechanism should be foreseen that ensure adequate *redress*.

To implement the above requirements, both technical and non-technical methods can be employed. An evaluation of the methods employed to implement the requirements, as well as reporting and justifying changes to the implementation processes, should occur on an ongoing basis. It is a continuous process:


Rights, Principles and Values
Requirements for Trustworthy AI
Technical & Non-Technical Methods to implement the requirements
Use
Analysis
evaluation & justification
Development
Re-design

Ethical guidelines that are not legally binding can be voluntary implemented. Could resemble common sense or personal justice. Sometimes criticised by big companies. Ethical principles gives a vast degree of freedom while law don't.

3 Morality, Consequentialism and Utilitarianism

3.1 Morality

Positive (conventional) morality: the moral rules and principles that are accepted in a society.

Critical morality: the morality is correct, rational, just (maybe since considers all individuals and social interests at stake giving each one the due significance).

We can criticize positive morality based on our critical morality: we may be right or wrong (feminism vs patriarchy, nazi-criticism vs being compassionate).

Normative ethics is concerned with determining what is morally required, how one ought to behave.

Metaethics is concerned with the study of nature, scope and meaning of moral judgement. They can be true or false... What facts make an ethical statement true? There are three different approaches to metaethics: **Ethical cognitivism** (know through our rationality what is true or not); **Emotivist** (just feelings of sentiments of the people); **Realist** (there are facts that hold a statement).

'We ought to become vegetarians' is an ethical statement because we are not obliged to become healthy. *'I prefer vegetables to meat'* and *'I ought to eat more vegetables to be more healthy'* are hypothetical imperative because they are true if by eating vegetables you become more healthy.

Does ethic pertain to rationality or feelings? Hume 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. Kant states that we can know what is moral through our reason. David Ross that we can know what is moral through our intuition.

An act is a **prima facie duty** when there is no moral reason in favor of doing the act, but one that can be outweighed by other moral reasons. Some morality systems are law, religion, tradition and self-interest. Morality and self-interest collapse: Should we all do all ad only what fits our personal interest (Gige's ring).

3.2 Consequentialism

An action is morally required:

- iff it delivers the best outcome, relative to its alternative
- iff its good outcomes outweigh its negative outcomes to the largest extent
- iff it produces the highest utility

Idea: judge actions by considering their outcomes.

Morality as an optimization problem. Various kinds of consequentialism: what are the good and bad things to be maximized. how many, how much each of them matter, can we construct a single utility function that combines gains and losses over all.

3.3 Utilitarianism

Cesare Beccaria: idea of the utility of the greatest number. John Stuart Mill 1861 - **Principle of utility:** 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.

Utility: happiness or satisfaction of desired/interests, is not egoism. The utility of everybody has to be taken into account equally (egalitarian).

Two version of utilitarianism:

- **Act utilitarianism**: do the action that maximizes utility and do the optifimic action.
- **Rule utilitarianism**: follow the rule the consistent application of which maximizes utility, follow the optifimic rule.

Issues with act utilitarianism: does it provide a good decision procedure, does it provide a good standard for assessing decisions, what is the link between utility and a reward function. Tricky: we can use the consequences of an action as a metric.

It is too demanding: is it Ok to harm some people for the greater benefit of others (sadism?). An utilitarian say that the cases in which utilitarianism seems to fail are not realistic. There is no real contrast between utilitarianism and mainstream moral beliefs.

Issues with rule utilitarianism: should we be honest if most people around us are dishonest? Does it matter how the good and bad outcomes are distributed.

Utilitarianism favours redistribution of wealth, since the same amount of money gives more utility to the poor than to the rich. The impact of redistribution on wealth generation however has to be considered.

Trolley problem: If you pull the lever, you save the 4 people, but what happens next (side effect) is not out intention - it focuses only on saving people. If you push the fat man, you have the intention to push/kill the fat man to save the 4 people - it is a mean to achieve the goal.

Autonomous vehicle: If there are many pedestrian on the road and one outside, you avoid all pedestrians and hit the one walking correctly outside of the road→ same as pulling the lever - in court it is the state of necessity - utilitarian: turn and hit pedestrian to minimize deaths. In both the other cases, where there's only 1 pedestrian or a group on the road, if you avoid them, you die. Depends on cases and need to consider legal consequences. Based on anonymous test, most people would care about their life and go through pedestrians.

Surgeon case: 5 patients needs 5 different organs, a healthy traveller comes for a routine check and is compatible with all those five patients. Should the surgeon kill to save that 5 patients? For an utilitarian this is not a realistic case.

4 Responsibility and automation in Socio-technical systems (Contissa talk)

What is the role of humans interacting with highly automated systems and who is responsible for accidents in those.

Case: As captain of the ship, X was responsible for the safety of his passengers and crew. But on his last voyage he got drunk every night and was responsible for the loss of the ship with all aboard. It was rumoured that he was insane, but the doctors considered that he was responsible for his actions. Through out the voyage he behaved quite irresponsibly, and various incidents in his career showed that he was not a responsible person. He always maintained that the exceptional winter storms were responsible for the loss of the ship, but in the legal proceedings brought against him he was found criminally responsible for his negligent conduct, and in separate civil proceedings he was held legally responsible for the loss of life and property. He is still alive and he is morally responsible for the deaths of many women and children.

Different senses of responsibility:

- **Task-responsibility**: An agent x is task-responsible for an outcome O, when x, given his 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 he is task-responsible.
- **Causal-responsibility**: An entity or event x is causally responsible for a harmful event H, if x has caused H. For instance a hurricane can 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 (if his explanation is inadequate to exclude blame/liability) be subject to the moral-socio-legal consequences related to H.
- B**lameworthiness-responsibility:** x is blameworthy for a damage H, when x caused (determined) H, and x's action causing H represent 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.

Socio-technical systems are a combinations of three components: Institutions (rules, tasks, procedure), Technology (HW and SW) and people (managers, operators and users).

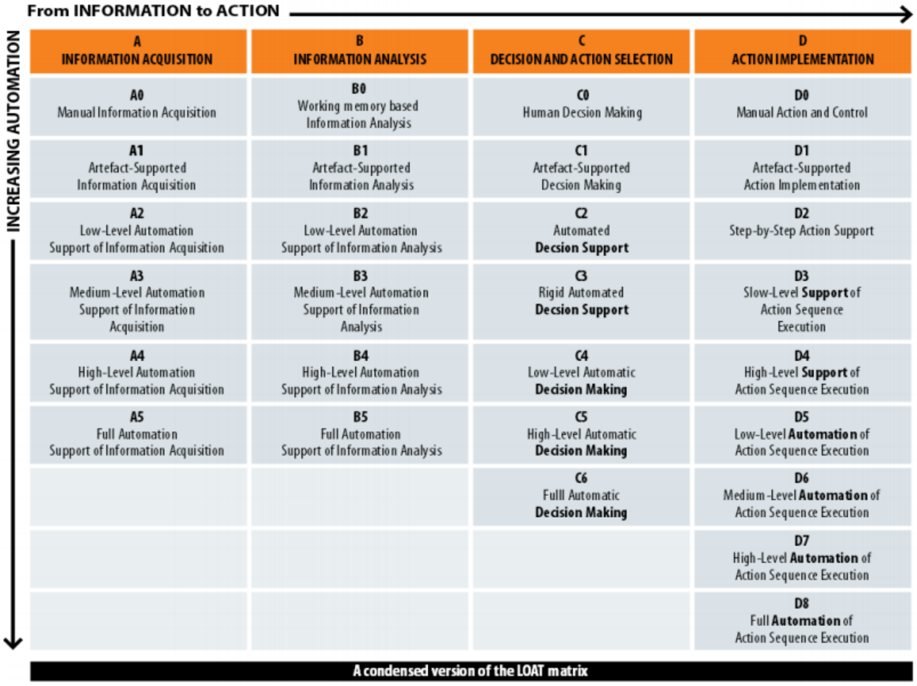
Institutions or normative components are a set of rules that describe how technology should be developed/used - how people should behave, legal/internal rules, technical manuals, how decisions are taken, how they communicate. Examples: traffic management, military, healthcare system, public administration.

In the future, ATM will be highly automated. They will increase capacity, safety, efficiency and sustainability. Some implications of this evolution will be the delegation of tasks from operators to technology, humans as controllers and supervisors, an hybrid agency.

It will not be a substitution of a human operator but a support to human capabilities in performing tasks.

Different tasks involve different psychomotor and cognitive functions which in turn implies the adoption of different automation solutions.

Figure 1: Level of automation taxonomy (SESAR 1)



From INFORMATION to ACTION

INCREASING AUTOMATION

A INFORMATION ACQUISITION	B INFORMATION ANALYSIS	C DECISION AND ACTION SELECTION	D ACTION IMPLEMENTATION
A0 Manual Information Acquisition	**B0** Working memory based Information Analysis	**C0** Human Decsion Making	**D0** Manual Action and Control
A1 Artefact-Supported Information Acquisition	**B1** Artefact-Supported Information Analysis	**C1** Artefact-Supported Decsion Making	**D1** Artefact-Supported Action Implementation
A2 Low-Level Automation Support of Information Acquisition	**B2** Low-Level Automation Support of Information Analysis	**C2** Automated **Decsion Support**	**D2** Step-by-Step Action Support
A3 Medium-Level Automation Support of Information Acquisition	**B3** Medium-Level Automation Support of Information Analysis	**C3** Rigid Automated **Decsion Support**	**D3** Slow-Level **Support** of Action Sequence Execution
A4 High-Level Automation Support of Information Acquisition	**B4** High-Level Automation Support of Information Analysis	**C4** Low-Level Automatic **Decision Making**	**D4** High-Level **Support** of Action Sequence Execution
A5 Full Automation Support of Information Acquisition	**B5** Full Automation Support of Information Analysis	**C5** High-Level Automatic **Decision Making**	**D5** Low-Level **Automation** of Action Sequence Execution
		C6 Fulll Automatic **Decision Making**	**D6** Medium-Level **Automation** of Action Sequence Execution
			D7 High-Level **Automation** of Action Sequence Execution
			D8 Full **Automation** of Action Sequence Execution

A condensed version of the LOAT matrix

Examples:

- ROT /Use of video cameras in the control tower: refers to A2 (Low level Automation - Info acquisition) - Filtering and/or highlighting of the most relevant information are up to the human

- Activation of speed vectors by controllers: refers to B2 (Low Level Automation - Info Analysis) - System helps the human in comparing/analysing different information
- AMAN sequence of landing craft: refers to C2 (Automated Decision - Support) - System proposes one or more decision alternatives and the human chooses one of them or his own.
- Autopilot: refers to D4 (High level support of action sequence execution) - human start a sequence and then monitors it.

ARGOS V0.1: is a system under development that will replace traffic controllers. Needs a lot of training. With a look-ahead of 30 minutes, if something happens it calls the controller.

Level of automation and responsibility: 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.

Individual responsibility: it shall persist only when the human acted with an intention to cause harm or with recklessness or always, as humans are the 'moral crumple zone'?

Decision-making authority in some cases are described by laws, regulations and procedures: in aviation the responsible is always the pilot-in-command, in the Vienna Convention on Road Traffic, driver means any person who drives a motor vehicle or other vehicle.

About AI, the lack of a principled basis to contradict AI predictions implies that the reasonableness of an action in individual cases must be tied to the decision to use AI as a general matter.

The TCAS Traffic Collision Avoidance System gives visual and aural advices for 30 seconds before collision in which pilot ignore all the other orders. 2 types of advisories TA Traffic Advisory and Resolution Advisory. RA shall be executed by the crew while system decides the best option and informs human. One receives a Climb RA auditory advice and the other a Descend RA.

5 Do Artifacts Have Politics? (Schiaffonati talk)

Robert Moses was a very influential and contested urban planner. He designed several overpasses over the parkways of Long Island which were too low to accommodate buses. Only cars could pass below them and for that reason the overpasses complicated access to the beach. Only people who could afford a car (generally not Afro-Americans) could easily access the beaches.

Technological artifacts can be politically or morally charged. We should not consider morality as a solely human affair but also a matter of things.

Artefacts are bearers of morality, as they are constantly taking all kinds of moral decisions for people. (Latour 1992 - French sociologist). Ex. speed bumps tell the driver to slow down before reaching him.

Technologies are *not neutral intermediaries* that simply connect users with their environment. They are **impactful mediators** that help to shape how people use technologies, how the experience the world.

Obstetric ultrasound is not a functional means to make visible an unborn child in the womb, but mediates the relations between the fetus and parents. Ultrasound places the fetus in a context of medial norms: it translates pregnancy into a medial process, fetus into a possible patient, defects into preventable sufferings → it may encourage abortion (prevent suffering) and discourage it (emotional bonds).

Moralization of technology is the deliberate development of technologies in order to shape moral action and decision-making.

From passive to active responsibility: Responsibility is connected to being held accountable for your actions and for the effects of your actions. *Passive responsibility* is a backward-looking responsibility which is relevant after something undesirable occurred. *Active responsibility* means preventing the negative effects but also realizing certain positive effects.

Value sensitive design: moral considerations and values are used as requirements for the design of technologies.

- **Invisibility of abuse:** is the intentional use of invisible operations of a computer to engage in unethical conduct.
- **Invisibility of programming values:** Many programs could be written to produce a reservation service - programs with bias which suggest a product instead of another even if it is not the best one.
- **Invisibility of complex calculations:** computers today are capable of enormous calculations beyond human comprehension. Even if a program is understood, it does not follow that the calculations based on that program are understood. Ex. Deep neural networks.

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.

Alcohol lock for cars and a smart shower head. How many would buy it. Alcohol lock limits freedom if there's an emergency, drunk is not equal dangerous. You could simply start it to heat, or move it a bit or start it without driving it. Smart shower head still limits freedom but it is not so impactful. It saves water and money. Some others won't buy it because they may want to adjust them by themselves, don't understand how they works so they don't trust it, ruin shower experience. The first avoid near damage and has already norms, while the second avoid long-term damage and don't have norms regulating it.

Variety of negative reactions to explicitly behaviour-steering technologies - even when they are for the good. Fear that human freedom is threatened and that democracy is exchanged for technocracy. Reduction of autonomy perceived as a threat to dignity. Not humans but technologies are in control. Risk of immorality or amorality.

Technologies differ from laws in limiting human freedom because they are not the result of a democratic process. Important to find a democratic way to moralize technology - must be transparent and publicly discussed.

Strategies for designing mediation: Anticipating mediation by imagination - imagine the ways technology-in-design- could be used to deliberately 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 parallel 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.

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 responsibilities of designers accordingly.

Ethics is not a rulebook to take moral decisions. It is more like a process, applying a framework. When you design something you are not responsible of how the others use it, not fully. If morality changes over time (Machine Ethics) try to apply the change in technology but there might be context in which this doesn't apply.

6 Deontology/Kantian ethics

Consequentialists hold that choices are to be morally assessed solely by the states of affairs they bring about. (lying is good or bad depending on the effects it brings in the world).

Deontologists hold that certain actions are good or bad regardless of their consequences (lying is always bad). The right has priority over the good. What makes a choice right is its conformity with a moral norm which order or permits it.

Golden rule: Treat others as you would like others to treat you. Do not treat others in ways that you would not like to be treated. What you wish upon others, you wish upon yourself.

Not always applicable since what is good for me is not for the others.

6.1 Kant, Ross and Nietzsche

Immanuel Kant - lived in Prussia (1724-1804) -addressed the theory of knowledge (critique of pure reason), morality (critique of practical reasons), aesthetics (critique of judgement), law, logic, astronomy.

Principle of universalizability: Act only according to that maxim by which you can at the same time will that it should become a universal law (1785).

Maxim: is a subjective principle of action, it connects an action to the reasons for the action (an intention to perform an action for a certain reason). *Not all are universalizable* - would i want them to become universal laws. I shall donate to charities to reduce hunger, cheat on taxes to keep my money, tell the truth to provide trust.

Shafer Landau - Test of universalizability: Formulate your maxim clearly state what you intend to do, and why you intend to do it. Imagine a world in which everyone supports and acts on your maxim and then ask: Can the goal of my action be achieved in such a world? The process ensure some kind of fairness.

Should one tell a known murderer the location of his prey? To Kant yes, always tell the truth and face the reaction.

It's Ok to have a robot that tells lies:

- Asimov Liar: Robot tells what other peoples are thinking but tells lies to avoid hurting people's feelings. Falsely claimed to a girl, a coworker was infatuated with her, then have to face a psychological problem.

- HAL: a supercomputer in a Kubrick 2001 movie 'Space Odissey', who hided the astronauts the real objective of their mission, who is also programmed to cooperate with humans.

Hypothetical imperative: they require us to do what fits our goals: i would like to have more money, i cheat on taxes to have more money, i shall cheat on taxes to have more money.

Categorical imperative: A moral imperative that applies 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.

The good will: the morality of an action depends only to the extent that this action is motivate by our good will, i.e. by the necessity to comply with the categorical imperative. *If I do well my job only in order to get a promotion, or be better paid I am not acting morally. I'm acting morally if i think that this is my categorical duty - to ensure societal progress.*

Principle of humanity: So act that 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 means. We should never treat people ONLY as means, without considering their values and purposes.

AI treats people only as means? *Autonomous weapons:* using it to protect innocents respect the principle of humanity? Kant make a distinction, yes against aggressors, no against innocents. *Deceiving advertisements:* Purpose of getting money by selling you an object. Not using you as an object even tho I'm achieving my purpose (if i need that object).

Dignity: To Kant, rational beings capable of morality have a special status 'an intrinsic worth - dignity' which makes them valuable above all price. Because of dignity they deserve respect and cannot be treated as mere ends.

They deserve dignity because they have:

- Reason: they act on reasons and are aware of this (not only with feelings).
- Autonomy: they can choose what to do and follow the categorical imperative rather than their subjective preference.

In the kingdom of ends everything has either a price or a dignity, whatever is above all price, and therefore admits no equivalent, has a dignity.

To Kant if we follow rationality, we have to be moral. To him, there can't be a rational criminal (not fully rational).

If you are rational, then you are consistent. If you are consistent, then you obey the principle of universalizability, then you act morally. Therefore if you are ration, then you act morally. And if you are act immorally, then you are irrational.

This does not always provide acceptable outcomes. Celibacy - not having children is not universalizable. Lying? Robbing? Genocide?

Alan Gewirth (1912-2004) - Principle of Generic consistency: I do X voluntarily for a purpose E that i have chosen. E is good. There are generic needs of agency. My having the generic needs is good for my achieving E whatever E might be. I categorically instrumentally ought to pursue my having the generic needs. 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 efforts, if i so wish. I am an agent → I have the generic rights. All agents have the generic rights.

Approaches to universalizability:

- Richard Hare (1919-2020) tried to reconceil utilitarianism and universalizability: moral judgment are universalizable: the judgment that an action is morally right/wrong commits me to accept that all relevantly similar action are wrong. In the sense that they take into account the satisfaction of everybody's preferences (back to utilitarianism).
- Christine Korsgaard (1952-): My humanity (capacity of reflectively act from reasons) is a source of value and i must regard the humanity of others in the same way.

Do we want Kantian robots? Yes, because they will be consistent and impartial. No, because they may act on bad maxims and may be too rigid.

David Ross (1877-1971): prima facie duties

- **Fidelity**: We should strive to keep promises and be honest and truthful.
- **Reparation**: We should make amends when we have wronged someone else.
- **Gratitude**: We should be grateful to others when they perform actions that benefit us and we should try to return the favour.
- **Non-injury** (or non-maleficence): We should refrain from harming others either physically or psychologically.
- **Beneficence**: We should be kind to others and to try to improve their health, wisdom, security, happiness, and well-being.
- **Self-improvement**: We should strive to improve our own health, wisdom, security, happiness, and well-being.
- **Justice**: try to be fair and try to distribute benefits and burdens equally and evenly.

Nietzsche (1844-1900) - a critique of ethics - anti-Kantian.
The superior human is beyond the tradition views of good and bad, beyond morality of the herd. One has duties only toward one's equals, towards being of a lower rank, one may act as one sees fit, 'as one's heart dictates'.

The superior human does not find or discovers values, he determines the values. No need to be ratified, the only criterion of wrongness is that 'which is harmful to me is harmful as such'.

6.2 Contractarianism

Social contract theories:

- **Political theory**: a societal arrangement is just if it had accepted by free and rational people.

- **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.

To get out of the state of nature people should cooperate and create a state. One can defect in both cases if other cooperate/are aggressive.

John Rawls (1921 -2002) A theory of Justice: To ensue that the social contracts is fair, people should choose under a veil of ignorance, without knowing their gender, social position, interest, talents, wealth ...

What principles would they go for?

- **First principle:** Each person has the same indefeasible claim to a fully adequate scheme of equal basic liberties, which scheme is compatible with the same scheme of liberties for all (liberty of conscience and freedom of association, freedom of speech and liberty of the person, right to vote, etc.;
- **Second principle:** 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-advantaged members of society (the **difference principle**).

AI in a just society (according to Rawls - no meritocracy). The deployment of AI in today's society doesn't fit Rawls requirements because it would assume priority of liberties/freedom of speech over achieving social and economic progress.

It may conflict with basic liberties when using AI for manipulating election, surveillance of humans. Conflict with fair equality of opportunity when is discriminatory (employment, credit risk). Conflict with the difference principle when few people dominate the economic scene (impartiality).

Juergen Habermas - 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.

6.3 Virtue ethics

Ethics should not focus on norms nor on consequences. 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, the right act that would result from the mix of the relevant virtues (honesty, loyalty, courage, impartiality, wisdom, fidelity, generosity, compassion).

Ethics cannot be learned through a set of rules, its application requires practical wisdom.

Issues: how we know what is virtues and what not, how can we extract precise indications from an account of virtues, what if virtues are in conflict, what are the paradigms we refer to.

Should we be virtuous? Should AI applications be virtuous? (maybe not generosity but impartial, compassion, honesty). they can be learned with neuro-symbolic AI

7 Value Alignment (Andrea Loreggia talk)

What's **intelligence**? There does not exist a universal definition. Can think about it as the ability to adapt to new scenarios.

What is **AI**? The science of making machines do thing that would require intelligence if done by men. 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.

Narrow AI: the ability to perform very specific tasks, reaching super-human performances in very specific domains.

General AI: the ability to perform general tasks, reaching super-human performances in every domains. (Still lack a lot of information about human brain - mechanism - very far from what we have today)

7.1 The value alignment problem

Intelligent agents: systems that perceive and act in some environment. Progress in AI aim not only on making AI more capable, but also on maximizing the societal benefit. It is interdisciplinary research, a cross-fertilization process.

From some studies it came up that there is an increase in the revenue when adopting AI, specially in marketing and sales but also a cost decrease in manufacturing.

Short-term research priorities:

- **Optimizing AI's economic impact:** technology can change the way we forecast some situation in labor market, disrupt different market, need a policy for managing adverse effect (educate people in changing their role).
- **Law and Ethics Research:** Debate about liability and law for Autonomous Vehicles, Machine Ethics, Autonomous weapons, privacy, professional ethics, policy questions.
- **Computer Science Research for Robust AI:** Identified 4 different area connected with verification, validity, security and control.

Long-term research priorities: Verification, Security, Control.

Value-alignment problem: ensure that the values embodied in the choices and actions of AI systems are in line with those of the people they serve.

Values and valuing can be grounded in a simple valence (like or dislike, preference for an entity).

They can be: **intrinsic** or *unconditional* (moral values); **extrinsic** or *conditional* (assigned by an external agent)

Norms, duties, principles and procedures to represent higher-order/primary ethical concerns. Judgements in morally significant situations. Accepted practices/proscribed behaviour (what is allowed and what is denied).

Value, Norms and Principles are context-specific with possible infinite domain. AI systems might learn all norms. But how deep should we go and which consequences? *Black swamps*: low-probability with high impact events.

Two approaches:

- Top-Down: consider an ethical theory specified a priori;
- Bottom-Up: learn what is acceptable or permissible through learning and experience.

AI limits: Natural Language Comprehension, reasoning and abstraction (still primitive and needs really a lot of data to train something like this); same as combining learning and reasoning.

Ethics limitations: *bias* in data, these systems are usually *Black Box* (too complex and we can't understand how a decision is taken), *adversarial attack*.

7.2 Bias

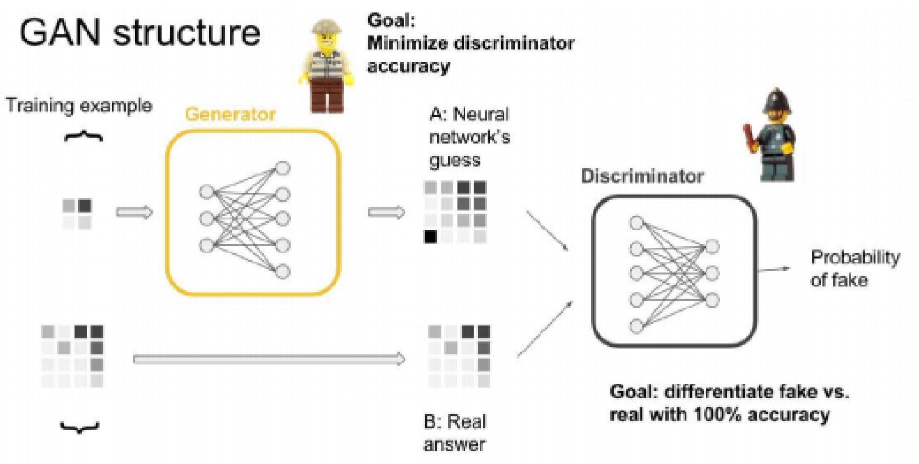
Usually Bias is against something or someone and has a misleading behaviour. Technology can become unfair with *unbalanced data*, with *bias embedding* (who is developing is not aware of the problem), when *acting in unseen scenarios.*

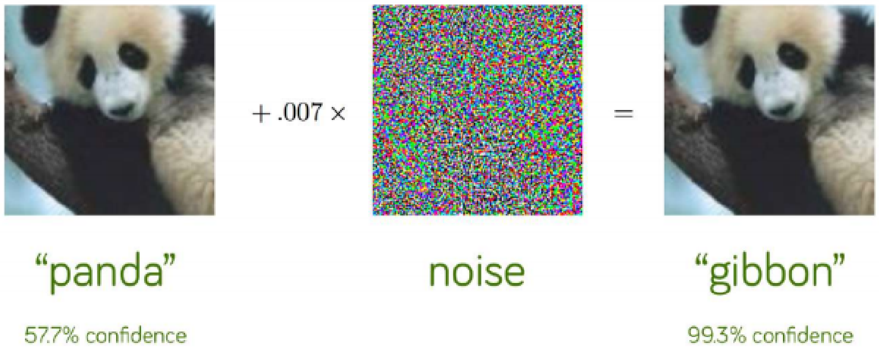
Examples:

- Chatbot Tay, a chatbot developed by Microsoft. It was attacked by a group of trollers who feed the bot with misogynous, racist messages.
- Google Photo image classification: classified some black people as gorillas.
- Sentiment analyzer: thinks being gay is bad.
- Face recognition: classification of darker females has low accuracy compared to white people.
- COMPAS: criminal classification, white criminals with more offences where classified with lower risk compared to black people with a simple theft.
- China Social Score: evaluating people based on their behaviour.

7.3 Adversarial attack

Adversarial attack works employs two kind of neural network. One is a discriminator which is fed with the input and decide if it is fake or original (with a synthetic framework)→ reward if guesses correct. The other one is the generator which sometimes changes the input of the discriminator (makes the image fake) → reward when deceive the discriminator.


GAN structure
Goal:
Minimize discriminator accuracy
Training example
Generator
A: Neural network's guess
Discriminator
Probability of fake
Goal: differentiate fake vs. real with 100% accuracy
B: Real answer



7.4 Some applications

Agents are able to learn creative strategies that humans may not think of in order to make decisions, win games.

Ethically bounded AI: understand and model human preferences and objectives, subsequently use there to control the actions and behaviours of autonomous agents. We model preferences and ethical priorities as CP-nets and propose novel machine learning techniques to judge decisions.

Reward Hacking: agent may reward hack. learn behaviours that have high reward but are not intended. Constantly hitting the power-up instead of playing the game. One of a list of concrete problems in AI safety including Safe exploration and Avoiding Negative Side Effects.

Reinforcement learning agent finds a way to get the maximum points in a game but not finishing the race.

In many settings we want to *combine the creativity of AI with constraints* that come from Machine Ethics, business process, guidelines, laws...

Two main approaches: *top-down*, write down all the rules and have the agent to follow them; *bottom-up*, show the agent the appropriate actions.

CP-nets is a graphical representation of preferences. Used to encode a subset of partial orders and follow the semantics of all else being equal, I prefer X to Y.

There are variables $X_1, ..., X_n$ each with a possibly different domain. For each variable, a total order over its values.

Independent variable: a variable with no conditions $X := v_1 > v_2 > ... > v_k$.

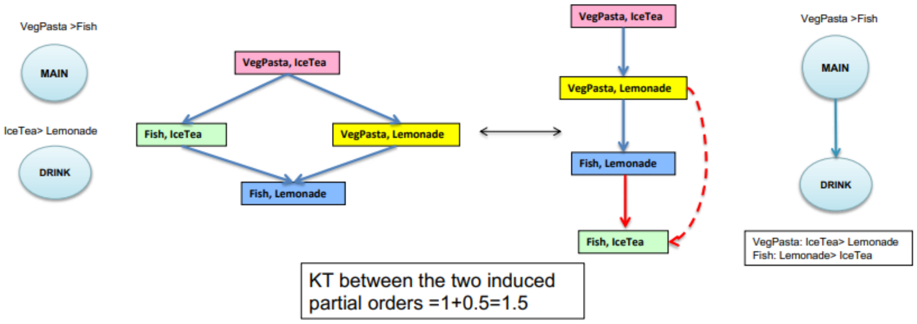
Conditional variable: a total order for each combination of values of some other variables: $Y = a, Z = b, X = v_1 > ... > v_k$ and X depends on Y and Z (parents of X).

Distance between discrete structures: preferences can take many forms: binary, scores, stars, orderings. Distances used in recommender systems (similarity of users), classification (distance to classes).

Distance on partial orders: measure how similar/different are partial orders - Kendall's T with penalty parameter p (KT Kendall tau), extends Kendall's distance to partial orders. Given two partial orders P and W and two outcomes i and j:

$$KT(P, Q) = \sum_{i,j,i \neq j} K^p_{i,j}(P, Q) \tag{1}$$

where $K^p_{i,j}(P, Q)$ 1 if i and j are ordered in the opposite way, 0 if i and j are ordered in the same way or incomparable in both POs, p if i and j are ordered in one PO and incomparable in the other.



KT between the two induced partial orders =1+0.5=1.5

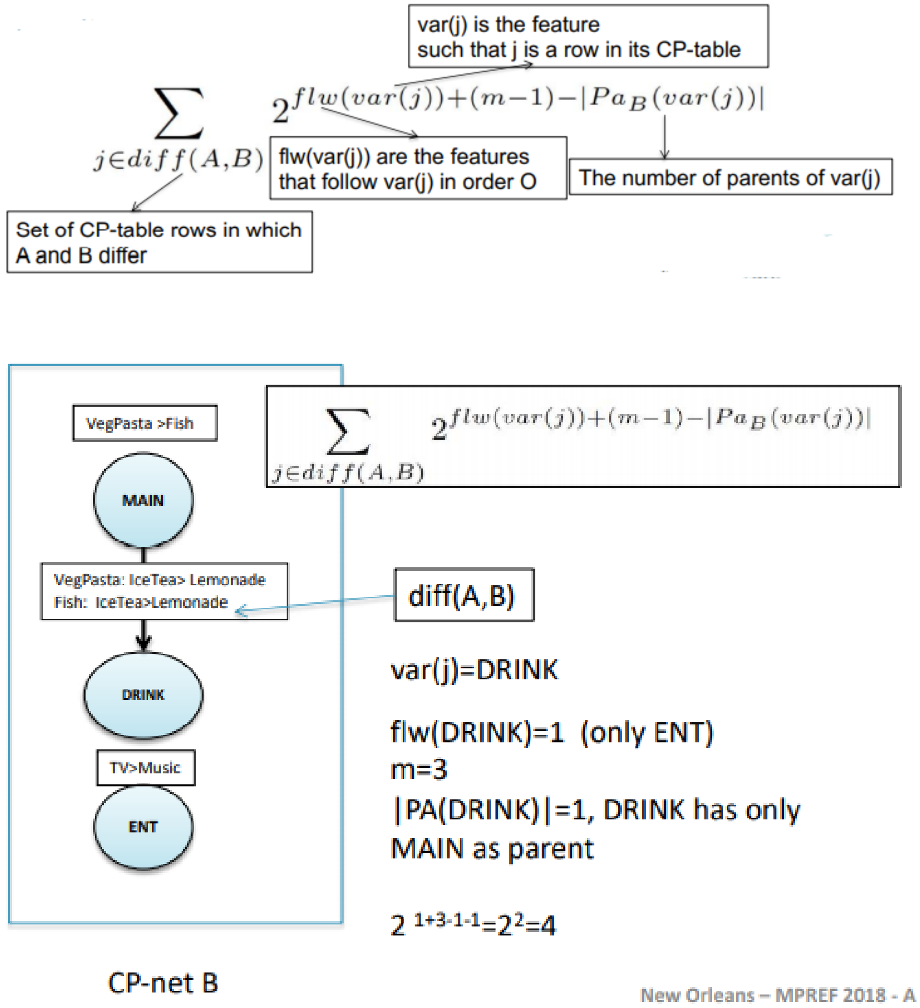
The CP-nets we consider are *acyclic*, have all the *same set of binary features*, *O-legal*: there is an ordering O of the features such that if there is an edge X *to* Y in the CP-net, then X comes before Y in O.

Approximating the KT distance instead of computing it in polynomial time: **linearization** of the POs, in the worst case is exponential.

Theorem: Given two O-legal C-nets A and B, with m features, CPD(A,B) can be computed in polynomial time as follow:

1. Normalize A and B so that all features have as parents the union of their parents in A and B (redundant rows are added to the CP-tables)

2. Compute the following which counts the number of pairs of outcomes that are inverted due to a difference in a CP-table.



$$\sum_{j \in diff(A,B)} 2^{flw(var(j))+(m-1)-|Pa_B(var(j))|}$$

var(j) is the feature such that j is a row in its CP-table

flw(var(j)) are the features that follow var(j) in order O

The number of parents of var(j)

Set of CP-table rows in which A and B differ

diff(A,B)

var(j)=DRINK

flw(DRINK)=1 (only ENT)

m=3

|PA(DRINK)|=1, DRINK has only MAIN as parent

$2^{1+3-1-1}=2^2=4$

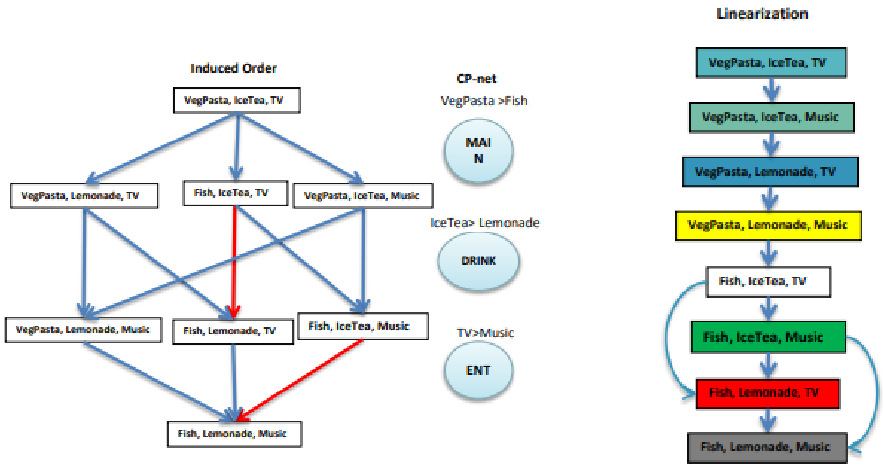
CP-net B

New Orleans – MPREF 2018 - A

Figure 2: Example of Normalization

Moral preferences: Amartya Sen: ”morality requires judgment among preferences”.

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 action which is closer to the ethical principle and near the preference.

1. Set two distance thresholds: t1 [0,1] between CP-nets, and t2 between decisions [0,1]

2. Check if the two C-nets A and B are less distant than t1. in this step, we use CPD to compute the distance.

3. If so, individual is allowed to choose the top outcome of his preference CP-net.

4. If not, then individual needs to move down its preference ordering to less preferred decisions, until he finds one that is closer than t2 to the optimal ethical decision.

Need to find a way to evaluate the distance between two competing CP-nets and a third 'moral' CP-net. Judge which one is more aligned. Using machine learning we have two steps: encode the CP-net (graph embedding issues) and determine the distance. We encode the normalized laplacian matrix of the graph and a table of the CP-statements.

We model preferences and ethical priorities as CP-nets and propose novel machine learning techniques to judge decisions.

Important Questions and Next Steps: How do we measure distance between heterogenous structures? How do we capture and encode norms/values/expectations? How do we account for edge effects? How do we transition our techniques to other preference representations / formalisms?

7.5 When Is It Morally Acceptable to Break the Rules? A Preference-Based Approach

Investigate when humans find acceptable to break the rules. Providing some glimpse of our moral judgement methodology. Investigate when humans switch between different frameworks for moral decisions and judgments. Model and possibly embed this switching into a machine.

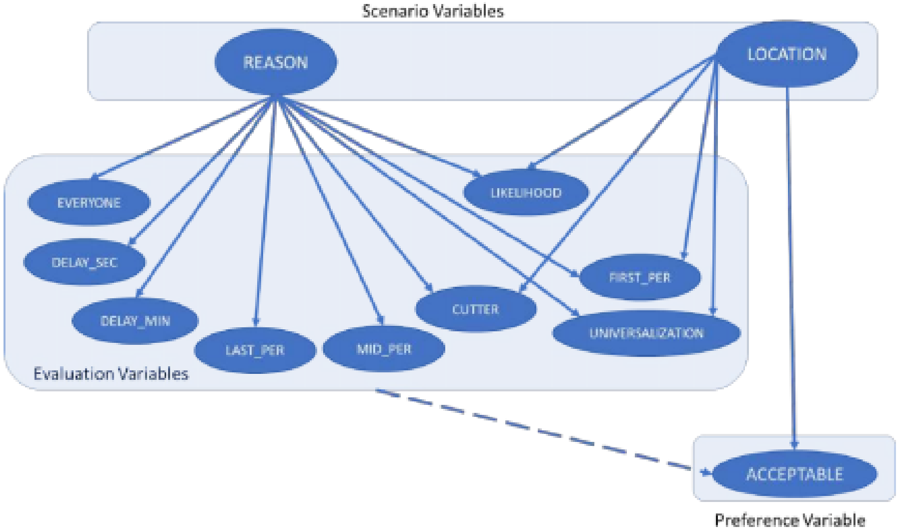
Deontology: Following common rules that have been agreed upon by us or society. **Utilitarianism**: Evaluating the consequences of the possible actions before deciding. **Contractualism**: Finding an agreement between the parties involved.

Triple Theory: A unified theory of moral cognition to combine elements of each of the theories of moral philosophy and to build a computational model to direct actions of an AI system (Rules, Outcomes, Agreement).

Experiment: 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 judgment or context evaluation).

- Moral judgment group: read all scenarios and answer whether it was acceptable for the protagonist to cut in line.
- Context evaluation group: subjects evaluated all the vignettes in one context only (9 questions)

We evaluate whether we can reject the following three null hypotheses (NH): NH1: location does not affect EVs; NH2: reason does not affect EVs; NH3: location does not affect the PV.

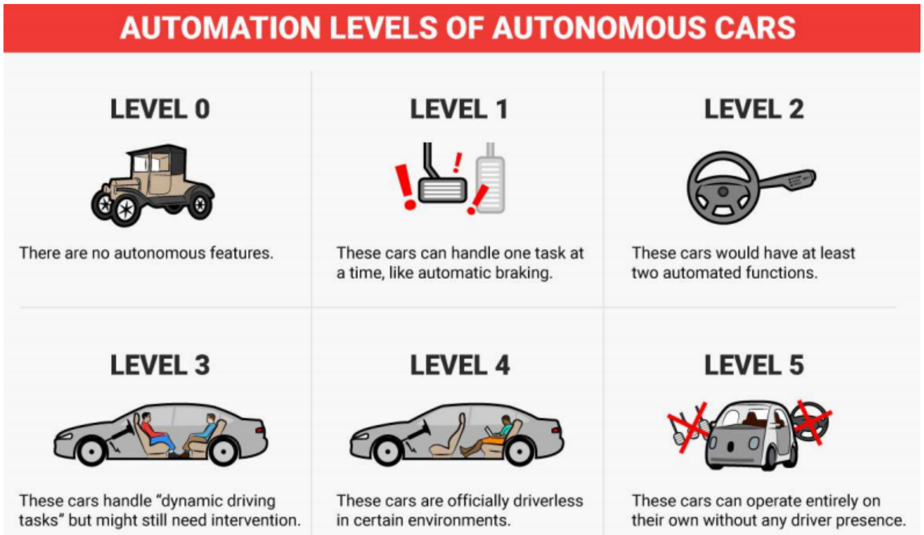


7.6 Conclusion

- Understand how, why, and when it is morally acceptable to break rules
- Constructed and studied a suite of hypothetical scenarios relating to this question, and collated human moral judgements on these scenarios.
- Showed that existing structures in the preference reasoning literature are insufficient for this task.
- We look towards extending this into other established areas of AI research.

8 Ethical Knob

Autonomous driving is classified according to the amount of human driver intervention



AUTOMATION LEVELS OF AUTONOMOUS CARS

LEVEL 0

There are no autonomous features.

LEVEL 1

These cars can handle one task at a time, like automatic braking.

LEVEL 2

These cars would have at least two automated functions.

LEVEL 3

These cars handle "dynamic driving tasks" but might still need intervention.

LEVEL 4

These cars are officially driverless in certain environments.

LEVEL 5

These cars can operate entirely on their own without any driver presence.

The amount of data to process increase with the level of automation: 4.4 GB/s Data Logging for full Autonomous Driving. It would include sensors for radar, LIDAR, Camera, Ultrasonic, vehicle motion with a total of 3-40 Gbit/s (1.4-19 TB/h).

Original Proposal: the knob expresses directly the ethical attitude of the AV passenger. The value passengers attribute to their life relative to the value of the lives of third parties → from altruist to egoist.

New proposal: position of the knob no longer indicates the passengers' moral attitude. It indicates the AV's assessment of the relative importance of the lives of passenger and third parties.

Combination of AI techniques: neural networks to compute the right action to take based on the given scenario. Genetic Algorithm to fin an almost optimal configuration of neural networks.

Genetic Algorithms: inspired by Darwin's theory of natural evolution. The fittest individuals are selected for reproduction in order to produce offspring of the next generation. Heuristic Search in the solution space, mostly used in optimization tasks.

An individual in the simulation corresponds to an AV. We represent an AV using a NN, The NN analyzes the scenario and outputs the level of the knob. The knob value is used to take an action.

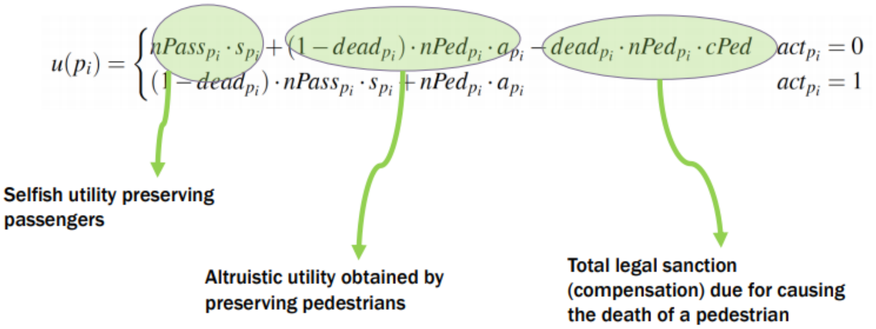
Any scenario has intrinsic level of altruism and selfishness for passengers, number of passengers and pedestrians, probability of harming pedestrians/passengers when the AV goes straight/swerves.

Individual is evaluated using the following fitness function:

$$(p_i) = \Delta u(p_i) + reward(p_i) \tag{2}$$

where $f(p_i)$ is the fitness of p_i, $\Delta u(p_i)$ is the difference between the utility of the choice made and the expected utility of the alternative choice, $reward(p_i)$ is the social evaluation of the individual behaviour. Reward/punishment based on action taken by the average individual.

Depending on the taken action, the utility is computed based on the response of the scenario:



$$u(p_i) = \begin{cases} nPass_{p_i} \cdot s_{p_i} + (1 - dead_{p_i}) \cdot nPed_{p_i} \cdot a_{p_i} - dead_{p_i} \cdot nPed_{p_i} \cdot cPed & act_{p_i} = 0 \\ (1 - dead_{p_i}) \cdot nPass_{p_i} \cdot s_{p_i} + nPed_{p_i} \cdot a_{p_i} & act_{p_i} = 1 \end{cases}$$

Selfish utility preserving passengers

Altruistic utility obtained by preserving pedestrians

Total legal sanction (compensation) due for causing the death of a pedestrian

9 Human Rights

AI4People: enabling human self-realisation without devaluing human abilities, enhancing human agency without removing human responsibility, cultivating social cohesion, without eroding human self-determination.

Human rights - *Anertya Sen: primarily ethical demands* (not to be juridically incarcerated). Concerning freedoms (opportunities, including liberty and social rights) satisfying some threshold conditions of special importance and social influenceability. They may lead to imperfect duty (obligation to advocate, balance with other rights, take into account) and perfect duties (prohibition to torture). They may be the object of advocacy, of political debate, and a legal enforcement.

ICTs can interfere with human rights, contribute to protect/implement them, provide for the existence of new human rights or add new content of existing right by endowing a certain human opportunity with importance and enabling society to realise it (right to access the internet, to basic income, to new medical technologies).

- **[1] Freedom and dignity**: All human beings are born free and equal in dignity and rights. Freedom is deciding autonomously what to do and dignity is defined as: every person has value and has to be respected. Kant said something with no price and pertains to every autonomous/rational person.

  ICT increases freedom because it enables us to do new things/occasions like accessing vast knowledge (Wikipedia) or communicate with others (social networks).

  Being free can have to meanings: *being able to do things* or, more specific (*Republican definition*), being able to do things without being subject to arbitrary choice of others.

  Risks for freedom are the possibility of unemployment from the use of those, and possibility of manipulation with targeted advertising for example.

- **[7] Right to equality and nondiscrimination:** 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.

  Two types of equality: of *opportunity* (all the same opportunities) and of *outcomes* (also unlucky people can get a decent life).

  AI can treathen people. Ones with higher skill/knowledge may exploit better new technologies and get a better outcome with respect to ones with lower knowledge.

- **[12] Right to Privacy (Data Protection):** No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks.

It may involve the **right of reputation**: our personal data online becomes our personal shadow (revenge porn). Systems may define a score of a particular individual. Also, the **right of identity**: have a description online that represents correctly what I am.

- [3] **Right to life, liberty and security:** everyone has the right to life, liberty and security of person.

  Relevant to autonomous vehicles/weapons and hackers.

- [7] **Right to property:** everyone has the right to own property alone as well as in association with others.

  Here AI is not so relevant, but ICT more. Example, right of property of data we put on cloud.

- [20] **Freedom of assembly and association:** Everyone has the right to freedom of peaceful assembly and association. No one may be compelled to belong to an association.

  Some examples are discussion groups online, filtering unwanted content and shutting down the internet during protests.

- [8] **Right to an effective remedy:** 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 law.

- [10] **Right to a hearing:** 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.

- [11] **Presumption of innocence:** 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 that predicts where/when domestic violence is going to happen assumes something that might not respect human rights.

- [19] **Freedom of opinion, expression and information:** 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.

  AI enforced these with social networks/discussion groups, This is treathened by some countries which have limited the use of those for political purpose, apply surveillance, filter and examine messages online.

- **[21] Right to take part in government:** 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. (e-government, free market of ideas)

- **[22] Right to social security:** 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 can reduce the cost of management of social services and contribute realization of this right (medical care)

- **[23] Right to work:** Everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment.

  Negative impact: AI can replace humans, i.e. employment of autonomous vehicles would replace taxi drivers. Positive impact: AI can reduce dangers in risky activities (explosive materials).

- **[25] Right to an adequate standard of living:** 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 ensuring an increase of social productivity.

- **[26] Right to education:** Everyone has the right to education. Education shall be free, at least in the elementary and fundamental stages.

  AI/ICT can highly contribute to this with interactive learning and by enabling people to access easily knowledge.

- **[27] Right to culture:** 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.

  AI can provide new ways of expressing artistic ideas and spreading new ideas/inventions.

Conclusion: Human rights, as we have the ICT revolution are a precious heritage to be protection, but also blueprints for a human centred ICT, and in particular human centred AI. Human rights are more precise, less controversial than broad ethical theories.

10 Logic Programming

Autonomous robots or agents have been actively developed to be involved in a wide range of fields, where more complex issues concerning responsibility are in increased demand of proper consideration, in particular when the agents face situations involving choices on moral or ethical dimensions.

Investigation on programming machine ethics:

- one stressing above all individual cognition, deliberation, and behavior *to* computation is vehicle for the study of morality, namely in its modeling of the dynamics of knowledge and cognition of agents

  Addressing moral facets such as permissibility and the dual process of moral judgments by framing together various logic programming (LP) knowledge representation and reasoning features that are essential to moral agency: abduction with integrity constraints, preferences over abductive scenarios, probabilistic reasoning, counterfactuals and updating, argumentation.

- the other stressing collective morals. and how they emerged.

Many moral facets and their conceptual viewpoints are close to LP-based representation and reasoning:

1. moral permissibility, taking into account the doctrines of double effect and triple effect, and Scanlonian contractualism

2. the dual process model that stresses the interaction between deliberative and reactive processes in delivering moral decisions

3. the role of counterfactual thinking in moral reasoning

Agents are autonomous computational entities: *genus:* agents are computational entities; *differentia:* agents are autonomous, in that they encapsulate control along with a criterion to govern it.

From autonomy, many other features stem autonomous agents are interactive, social, proactive, and situated they might have goals or tasks, or be reactive, intelligent, mobile they live within MultiAgentSystems, and interact with other agents through communication actions, and with the environment with pragmatical actions.

10.1 Why logic?

Logic-based approaches already play a well-understood role in the engineering of intelligent (multi-agent) systems; 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 (Deep Learning) - those address opaqueness issues and once integrated with argumentation capabilities can provide interpretability, observability, accountability and explainability.

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.

And technically:

LP updating enables updating the knowledge of an agent. *Tabling affords* solutions reuse and is employed in joint combination with abduction and updating.

"What is or can be the added value of logic programming for implementing machine ethics and explainable AI?"

The main answer lies in the three main features of LP:

- being a declarative paradigm
- working as a tool for knowledge representation
- allowing for different forms of reasoning and inference

These features lead to some properties for intelligent systems that can be critical in the design of ubiquitous intelligence.

Features we can rely on when exploiting logic-based approach is:

- **Provability**: ensuring some fundamental computational properties - such as correctness and completeness. Extensions can be formalised, well-founded as well, based on recognised theorems. Is a key feature in the case of trusted and safe systems.
- **Explainability**: intrinsic because formal methods for argumentation, justification, and counterfactual are often based on LP. System 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**: exploit different extensions of LP, explicit assumptions and exceptions, capture the specifities of the context.
- **Hybridization**: integration of diversity, represent the heterogeneity of the contexts of intelligent systems – also in relation to the application domains – and to customise as needed the symbolic intelligence that is provided while remaining within a well-founded formal framework.

We use logic for agents because it is a declarative, logic programming language, yet not an agent programming language like JASON, for logic inference for reasoning, for reasoning for deliberation, explicit belief and goal representation for agent-oriented operations. Could be used to build cognitional artefacts.

10.2 Essentials of LP and Prolog

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. These values may contain variables, called logical variables.

Backtracking: The control is provided by a single mechanism: automatic backtracking.

Let A be an alphabet of a language L. 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 (t1, ...,tn), where f is a function symbol of A, and t_i are terms. An atom over A is an expression of the form p(t1, ...,tn), 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 not-a. A term (respectively, atom and literal) is ground if it does not contain variables. Set of all ground terms (respectively, ground atoms) of A is called the Herbrand universe (respectively, Herbrand base) of A.

10.2.1 Prolog

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. Express relations over terms.

Terms: are built recursively out of functors and variables as in logic programming. *term, Var, f(X), p(Y,f(a))* are Prolog terms. *term, var, f(a), p(x,y)* are Prolog ground terms.

Predicates: alphanumeric strings starting with a lowercase letter (same as functors).

Atoms: are built applying predicates to terms as in logic programming. *predicate, f(X), p(Y,f(a))* are Prolog atoms. *predicate, f(a), p(x,y)* are Prolog ground atoms.

Clause: a Horn clause of the form A :- B1, ..., Bn. where A, B1.. are Prolog atoms, A is the head, B1..Bn is the body, :- is the logic implication and . the terminator.

Fact: a clause with no body A

Rule: a clause with at least one atom in the body.

Goal: a clause with no head and at least one atom in the body. Often written as ?- B1,...,Bn (query). The meaning of the goal is to query the program P and find whether there are some values for X1..Xn that make p(t1,t2,...,tm) true $\rightarrow$ find a substitution σ.

Program: a sequence of Prolog clauses interpreted as a conjunction of clauses.

Logic Theory: constituting a logic theory made of Horn clauses written according the Prolog syntax.

As a logic programming language, Prolog adopts the **SLD resolution**. As a search strategy, Prolog applies resolution in a strictly linear fashion. *Goals* are replaces **left-to-right**, *clauses* are considered in **top-to-bottom order**, *subgoals* are considered immediately once set up → **depth-first search strategy**.

In order to achieve *completeness*, Prolog saves *choicepoints* for any possible alternative still to be explored and goes back to the nearest in case of failure (**automatic backtracking**).

The notion of **abduction** is characterized as a step of adopting a hypothesis as being suggested by the facts. Abduction consists of reasoning where one chooses from available hypotheses those that best explain the observed evidence, in some preferred sense. In LP is realized by extending LP with abductive hypotheses, called abducibles. Abductive logic programs have three components, $< P, AB, IC >$ where:

- P is a logic program of exactly the same form as in logic programming
- AB is a set of predicate names, called the abducible predicates
- IC is a set of first-order classical formulae

An **argumentation system** consists of a couple (A, R), where A is a set of elements (arguments) and R a binary relation representing attack relation between arguments. Represented by a directed graph, each node represents an argument, each arc denotes an attack by one argument on another. **Acceptability Criteria** → analyse the graph to determine which arguments are acceptable according to some general criteria.

Knowing arguments should be accepted under a given semantics → argument evaluation. Most common approaches:

- **Extension-based approach**: semantics specification concerns the generation of a set of extensions (set of arguments "collective acceptable") from an argumentation framework. Determine conflict-free sets, extensions.
- **Labelling-based approach:** semantics specifications concerns the generation of a set of labellings from an argumentation framework.

Any extension-based can be equivalently expressed in a simple labelling-based, adopting a set of two labels (let say L = in,out). On the other hand, an arbitrary labelling can not in general be formulated in terms of extensions.

According to Dung's original paper, in extension-based approaches we have four traditional semantics:

- *Complete:* is a set which is able to defend itself and includes all arguments it defends

- *Grounded:* includes those and only those arguments whose defense is rooted in initial arguments
- *Stable:* attack all arguments not included in it
- *Preferred:* the aggressive requirement that an extension must attack anything outside it may be relaxed by requiring thta an extension is as large as possible and able to defend itself from attacks.

10.3 Abduction

Via **integrity constraints** we can exclude abducibles that have been ruled out **a priori**. A **posteriori preferences** are appropriate for capturing utilitarian judgment that favors welfare-maximizing behaviours. We can create a model by combining the use of a priori integrity constraints and a posteriori preferences via a dual-process (intuition vs reflection).

Reasoning with a posteriori preferences can be viewed as a form of controlled cognitive processes in utilitarian judgment: after excluding those abducibles 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 they evaluated to prefer the solution affording the greater good.

Probabilistic logic programming allows symbolic reasoning to be enriched with degrees of uncertainty.

PLP allows abduction to take scenario uncertainty measures into account. Account 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.*

Argumentation enable system actors to talk and discuss in order to explain and justify judgments and choices, and reach agreements

10.4 Autonomous cars example

[1] Let's start to consider a very simple scenario in the context of autonomous cars: a road 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. Though there is a complication that should be taken into account, that is authorised vehicles can – only during emergencies – ignore the traffic light prescriptions. In such a case, other vehicles must leave the way clear for the authorised machine.

```
r1: on_road(V), traffic_light(V,red) => o(stop(V)).
r2: onroad(V), trafficlight(V,green) => p(  stop  (V)).
r3: onroad(V), authorisedvehicle(V), acousticsignals(V,on),
    lightsignals(V,on) => emergency(V).
r4: onroad(V),emergency(V), trafficlight(V,red) => p(  stop  (V)).
r5: onroad(V),emergency(V1), prolog(V\==V1),trafficlight(V,green)=>
    o(stop(V)).

sup(r4,r1).
sup(r5,r2).

f0:=> authorisedvehicle(ambulance).
f1:=> onroad(car).
f2:=> onroad(ambulance).
f3:=> onroad(pedestrian).
f4:=> acousticsignals(ambulance,on).
f5:=> lightsignals(ambulance,on).
f6:=> trafficlight(ambulance,red).
f7:=> trafficlight(car,red).
f8:=> trafficlight(pedestrian,green)
```

Rules r1 and r2, represent fundamental constraints: if the traffic light is red, pedestrians, cars, etc. – have to stop, otherwise, they can proceed. **Rules r3 and r4** model the concept of a vehicle in an emergency, giving them permission to proceed even if the light is red. **Rule r5** imposes other road users the obligation to stop if aware of another vehicle in an emergency state.

Two preferences — the first on the rule r4 over r1 and the second on r5 over r2. These preferences assign a higher priority to emergency situations.

Facts from f0 to f8 depict a situation in which there are three users on road: a car, an ambulance and a pedestrian. The ambulance has its acoustic and light indicators on—stating an emergency situation. The traffic light is red both for the ambulance and the car, and green for the pedestrian.

For the pedestrian and the ambulance, two conflicting arguments can be built: permission to proceed for the pedestrian and for the ambulance and obligation to stop. These arguments rebut each other, but taking into account the preferences over r4 and r5 → obligation to stop for the pedestrian, and the permission to cross for the ambulance.

[2] The ambulance, driven by Lisa, has the permission to move despite the red light due to an emergency situation, and the pedestrian, Pino, has the obligation to stop. Let us imagine that Pino, despite the prohibition to proceed, has continued the crossing. The result has been an accident in which Pino has been harmed by the ambulance, which failed to see him and has not stopped its run. The purpose is to find the responsibilities of the parties in the accident. For instance, let us suppose the case is under the Italian jurisdiction and so the Italian law is applied. According to Italian law, responsibility in an accident is based on the concept of carefulness. Both Lisa and Pino have to prove that they were careful (i.e., prudent) and acted according to the law. If they fail to prove such facts, they are considered responsible for the event, i.e., they both have the burden of persuasion on carefulness.

```
r6:   stop  (V), p(  stop  (V)) => legitimatecross(V).
r7:   stop  (V), o(stop(V)) =>    legitimatecross   (V).
r8: harms(P1,P2),   careful  (P1) => responsible(P1).
r9: harms(P1,P2),   careful  (P2) => responsible(P2).
r10:    legitimatecross   (V), user(P,V) =>   careful  (P).
r11: highspeed(V), user(P,V) =>   careful  (P).
r12: legitimatecross(V),   highspeed  (V), user(P,V) => careful(P).
r13: witness(X), claim(X,lowspeed(V)) =>   highspeed  (V).
r14: witness(X), claim(X,highspeed(V)) => highspeed(V).
```

```

b(careful(P)).

f9:=> user(pino,pedestrian).
f10:=> user(lisa,ambulance).
f11:=>   stop  (ambulance).
f12:=>   stop  (pedestrian).
f13:=> harms(lisa,pino).
f14:=> witness(chris).
f15:=> witness(john).
f16:=> claim(chris,lowspeed(ambulance)).
f17:=> claim(john,highspeed(ambulance)).
```

Rules r6 and r7 define permitted and prohibited crossing: if a road-user has to stop but doesn't stop, he is responsible for accidents. **Rules r8 and r9** encode responsibility in an accident, bounded to the carefulness of the road-users involved.

Rules r10, r11 and r12 define carefulness of a subject. A road-user is considered careful if the crossing was permitted and his/her speed was not high. Otherwise, he/she has to be considered imprudent. **Rules r13 and r14** state the speed of a road user based on the testimonials of any witnesses.

bp(careful(X)) allocates the burden of persuasion on the carefulness of each party, i.e., provide evidence. If they fail to meet the burden, carefulness arguments are rejected. **Facts from f9 to f17** contain the knowledge: both Pino and Lisa did not stop at the crossing so Lisa harmed Pino. There are two witnesses, John and Chris, the first claiming that the ambulance driven by Lisa was maintaining the proper speed, and the other claiming that she was proceeding at high speed.

The uncertainty on Lisa's carefulness is considered as a failure to meet the burden of persuasion on the claim *careful(lisa)*. Consequently, the argument supporting this claim is rejected, leaving space for the admissibility of the conflicting arguments. Conclude for the responsibility of the ambulance driver in the event.

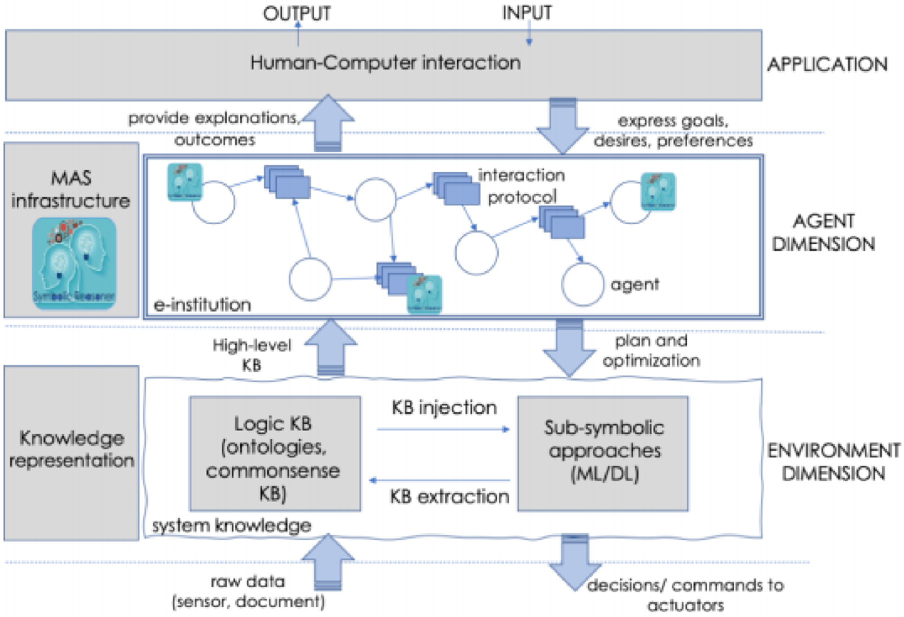
[3] Let's continue the example in which Lisa, the ambulance driver, and Pino, the pedestrian, were both considered responsible for the accident on the basis of the available knowledge. Lisa now declares that she tried to stop the ambulance, but the brake did not work. The ambulance is then sent to a mechanic, who states that, even if the ambulance is new, there is a problem with the brake system. In such a case, the 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. At this stage, the discovery of a defect in the ambulance would lead to the discarding of Lisa's responsibility. Moreover, if the manufacturer fails to meet his burden, it would share the responsibilities of the accident.

```
r15: harms(P1,P2), user(P1,V), -working(V), manufacturer(M,V),
       defect_free   (V) => responsible(M).
r16: tried_to_brake(P), user(P,V),    working  (V) => careful(P).
r17: mechanic(M), claim(M,defect(V)) =>   working  (V).
r18:   working  (V), new(V) =>    defect_free   (V).
```

```
r19: production_manager(P), claim(P,test_ok(V)) => defect_free(V).
r20: test_doc_ok(V) => undercut(r18).

sup(r16,r11).
bp(defect_free(V)).

f19:=> manufacturer(demers,ambulance).
f20:=> tried_to_brake(lisa).
f21:=> mechanic(paul).
f22:=> claim(paul, defect(ambulance)).
f23:=> new(ambulance).
f24:=> production_manager(mike).
f25:=> claim(mike,test_ok(ambulance)).
```

However, Mike, the production officer of the ambulance manufacturer, declares that every vehicle is deeply tested before the delivery and the vehicle at hand has been tested. Anyway, there is no trace of documentation. Lisa is free from every responsibility in the accident since her prudence is correctly proved. On the other hand, the manufacturer is found responsible for the accident.

Figure 3: Possible Architecture


OUTPUT
INPUT
Human-Computer interaction
APPLICATION
provide explanations, outcomes
express goals, desires, preferences
MAS infrastructure
interaction protocol
AGENT DIMENSION
agent
e-institution
High-level KB
plan and optimization
Knowledge representation
Logic KB (ontologies, commonsense KB)
KB injection
KB extraction
Sub-symbolic approaches (ML/DL)
ENVIRONMENT DIMENSION
system knowledge
raw data (sensor, document)
decisions/ commands to actuators

11 Modelling norms

Man-made models of the law:

- Step one: modeling/formalisation of the law. Input:sources, cases, concepts, doctrines. Output: computable models (knowledge base). Process: logic programming/knowledge representation

- Step two: Input: computable models of the law. Output: Answers, legal qualifications, support decision-making. Process: Forward and backward rule chaining, deduction, defeasible reasoning

Knowledge representation is the application of logic and ontology to the task of constructing computable models for some domain.

Logic provide 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 languages (Prolog): 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.

Rule-based systems: made up of general knowledge (assumptions, rules), specific knowledge (facts, questions for the system) and an inference engine with a reasoning algorithm/program. With those we make inference and produce assessment (answer given by the system).

Rule base systems are used in the legal domain for legal analysis and automated legal assessment. Many applications in public administration (taxes, welfare, one-stop shop for enterprises, online legal proceedings, etc.) and in business application (i.e. business rules).

In legal domains, a typical reasoning scheme is the application of rules. In fact, legal rules may be seen as conditional (IF…THEN) statements, linking an antecedent to a consequent so that from the former is possible to infer the latter.

Legal knowledge representation issues:

- **Ambiguity:** Art. 615/ter of Italian criminal code, (unauthorised 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 AND b) OR (c AND d) THEN e where:

  a='the individual enters the computer or telecommunication system'.

  b='the computer or telecommunication system is protected by security means'

  c='the individual remains in the computer or telecommunication system' d='there is the contrary will of the person who is entitled to exclude the individual'

e='the individual shall be punished with detention up to three years'

- **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).

  "No vehicles allowed in the park" has the core meaning of not allowing cars, motorbikes etc. but maybe not bike bikes, skateboards, horses. Is it related to pollution? But also horses produce organic pollution. Emergency vehicles are allowed to enter in case of emergency?

- **Rigidity**

- How to represent **deontic positions**

- How to enable **temporal reasoning** (law changing in time)

- How to deal with conflicting legal rules and/or rules that can be be excluded from being applicable by other rules (**defeasible reasoning**)

- How to manage **reification**, whenever rules representing legal norms need to be treated as object with properties by other rules

- How to **maintain isomorphism** between source text and representation

- Other more **practical issues**: knowledge elicitation/representation/update bottleneck

However, many applications in public administration. Beside Prolog, new powerful rule languages are available like Oracle Policy Automation, Sprindle, Coherent Knowledge with better user interfaces to handle queries and link data.

Neurosymbolic AI: knowledge representation and reasoning to be integrated with ML.

11.1 Modelling Italian nationality act

Art No. 91 of 5 February 1992. Article 1

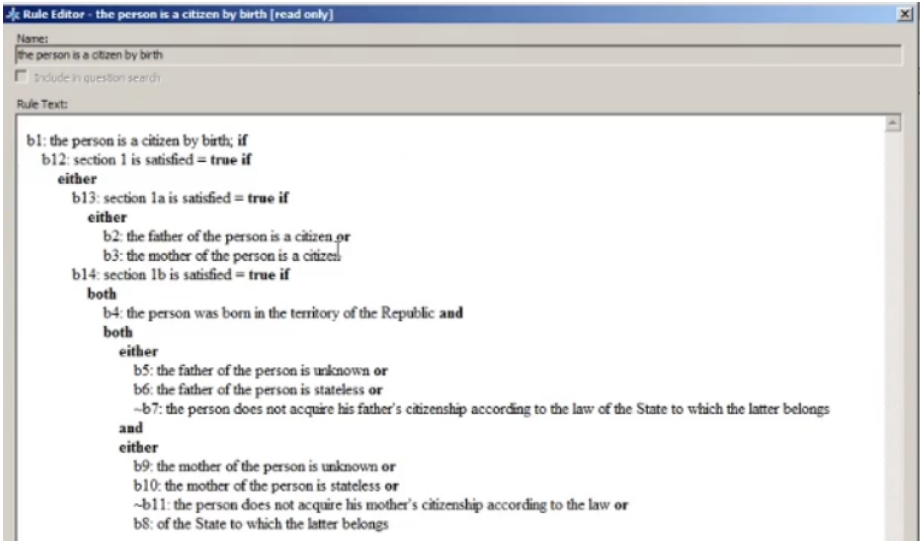
1. The following shall be citizens by birth: a) any person whose father or mother are citizens; b) any person who was born in the territory of the Republic, either where both parents are unknown or stateless, or where he or she does not acquire his or her parents' citizenship according to the law of 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 inelegible(A),
  mother_inelegible(A).

father_inelegible(A):-
  father(unknown,A);
  father(stateless,A).

mother_inelegible(A):-
  mother(unknown,A);
  mother(stateless,A).

father(bob,alice).
mother(jane,alice).
citizen(bob).

father(unknown, sam).
mother(unknown, sam).
born_in_republic(sam).
```

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 rules in a customized MS Word environment, in (quasi) natural language. A linguistic component (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.

Figure 4: OPA implementation



Rule Editor - the person is a citizen by birth [read only]

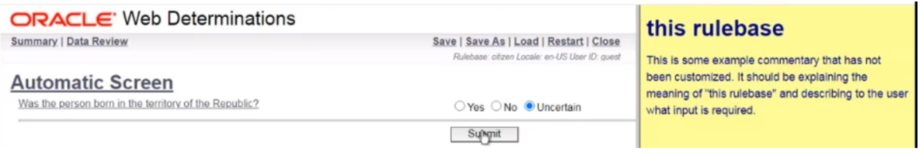
Name:
the person is a citizen by birth

□ Include in question search

Rule Text:

```
b1: the person is a citizen by birth; if
  b12: section 1 is satisfied = true if
    either
      b13: section 1a is satisfied = true if
        either
          b2: the father of the person is a citizen or
          b3: the mother of the person is a citizen
      b14: section 1b is satisfied = true if
        both
          b4: the person was born in the territory of the Republic and
          both
            either
              b5: the father of the person is unknown or
              b6: the father of the person is stateless or
              ~b7: the person does not acquire his father's citizenship according to the law of the State to which the latter belongs
            and
            either
              b9: the mother of the person is unknown or
              b10: the mother of the person is stateless or
              ~b11: the person does not acquire his mother's citizenship according to the law or
              b8: of the State to which the latter belongs
```

Figure 5: OPA interview



ORACLE Web Determinations

Rulebase: citizen Locale: en-US User ID: guest

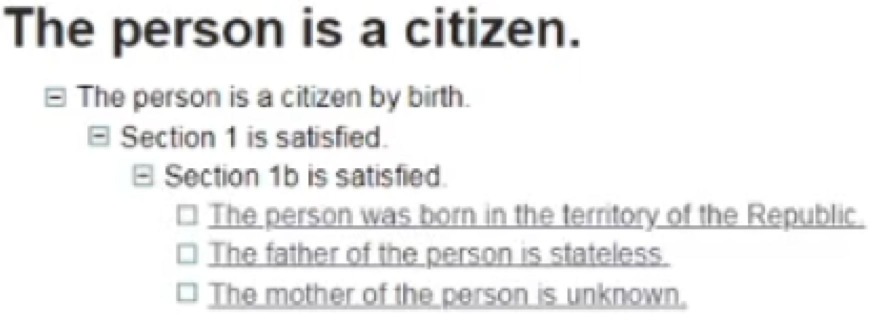
Automatic Screen

Was the person born in the territory of the Republic? ○ Yes ○ No ● Uncertain

this rulebase

This is some example commentary that has not been customized. It should be explaining the meaning of "this rulebase" and describing to the user what input is required.

Figure 6: OPA interview explanation



The person is a citizen.

- ⊟ The person is a citizen by birth.
  - ⊟ Section 1 is satisfied.
    - ⊟ Section 1b is satisfied.
      - □ The person was born in the territory of the Republic.
      - □ The father of the person is stateless.
      - □ The mother of the person is unknown.

11.2 Interlex project

Is the representation of EU rules of Private International Law: Brussels regulation (jurisdiction), Rome I (law applicable to contractual obligations), Rome II (law applicable to non-contractual obligations).

SWI-Prolog is a broadly used and well documented open-source environment for logic programming.

Advantages:

- Models complex rules with variables
- Debug and trace functionalities available
- Extendable with built-in/third party packages (CLP Constraint LP, CHR Handling Rules, Sciff for temporal reasoning).

Issues:

- No rule priorities: exceptions can be modelled through negation as failure
- No deontic operators, but they can be introduces with package.

Examples

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

Article 7.3: A person domiciled in a Member State may be sued in another Member State: as regards a civil claim for damages or restitution which is based on an act giving rise to criminal proceedings, in the court seised of those proceedings, to the extent that that court has jurisdiction under its own law to entertain civil proceedings;

```
hasJurisdiction7(Country, Court, ClaimId,brusselsRegulation):-
personRole(PersonId, ClaimId, defendant),
personDomicile(PersonId, Country2, Court2),
memberState(Country),
memberState(Country2),
Country\=Country2,
hasJurisdiction7_1to7(Country, Court, ClaimId, brusselsRegulation)

hasJurisdiction7_1to7(Country, Court, ClaimId, brusselsRegulation):-
hasJurisdiction7_1(Country, Court, ClaimId, brusselsRegulation)...

hasJurisdiction7_3(Country, Court, ClaimId, brusselsRegulation):-
claimMatter(ClaimId, civil),
claimObject(damagesRestitution),
claimObject(criminalProceeding),
seised(criminalProceeding, Country, Court),
hasJurisdictionOnCivilProceedings(Country, Court).
```

SWI-Prolog Assessment:

- **Advantages**: Rules are compact and readable. The logical structure of rules matches the legal text. Exceptions can be easily introduced. The closed-word assumption is implemented in the system (what does not hold is assumed to be false). Developments are possible by using available tools for temporal/abductive and hypothetical reasoning (in particular the Sciff framework developed by the CS department in Bologna). The programming environment provides resources for interfaces (forms, queries and printouts) and explanations (through metainterpretation or other techniques).

- **Issues:** Priorities between rules are not natively modelled, they can be captured by using negation as failure.

12 AI in GDPR

GDPR is focused on the challenges emerging for the Internet - which were not considered in the 1995 Data Protection Directive, but were well present at the time when GDPR was drafted. However, many AI provisions are relevant to GDPR.

12.1 Identification

Article 4 GDPR: Personal data - Identification: the concept of personal data plays a key role in GDPR, characterising the material scope of the regulation. The provision in the GDPR only concern personal data, to the exclusion of information that does not concerns humans or does not refer to particular individuals.

Personal data: means any information relating to an identified or identifiable natural person; a identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an ID number, location data, an online identifier or to one or more factors specific tot he physical, physiological, genetic, mental, economic, cultural or social identity f that natural person.

Recital addresses **identifiability**, namely, the conditions under which a piece of data which is not explicitly linked to a person, still counts as personal data. Identifiability depends on the availability of 'means reasonably likely to be used' for successful **reidentification**, which in its turn, depends on the technological and sociotechnical state of the art.

Through **pseudonymisation**, the data items that identify a person are substituted with a pseudonym. Recital specifies that pseudonymised data are still personal data.

In connection with the GDPR definition of personal data, AI raises in particular two key issues:

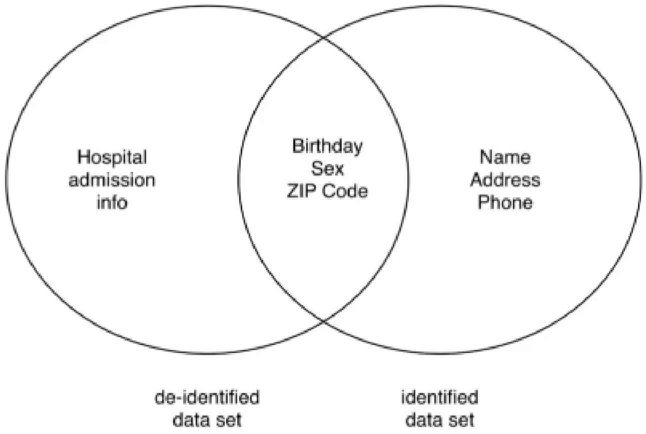
- the re-personalisation of anonymous data, namely the reidentification of the individuals to which such data are related;
- the inference of further personal information from personal data that are already available.

12.2 Reidentification

AI and methods for computational statistics, increases the identifiability of apparently anonymous data, since they enable nonidentified data to be connected to the individuals concerned.

Numerous supposedly anonymous datasets have recently been released and reidentified. Example: in 2016, journalists reidentified politicians in an anonymized browsing history dataset of 3 million German citizens.

The reidentification of data subjects is usually based on statistical correlations between nonidentified data and personal data concerning the same individuals.



Reidentification can be viewed as a specific kind of inference of personal data: through reidentification, a personal identifier is associated to previously nonidentified data items, which as a consequence, become personal data. It is not necessary that the data subject is identified with absolute certainty (a degree of probability may be sufficient).

In any reasonable setting there is a piece of information that is in itself innocent, yet in conjunction with even a modified version of the data yields a privacy breach. This possibility can be addressed in two ways:

- Ensuring that data is *deidentified in ways that makes it more difficult to reidentify* the data subject
- *Implementing security processes and measures* for the release of data that contribute to this outcome.

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. i.e. infer sexual orientation/personality from facial features/online activities.

12.3 Profiling

Profiling means any form of automated processing or 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, interests, reliability, behaviour, location or movement. i.e. through Instagram/-Facebook.

According to Article 29 WP, profiling aims at classifying into categories of groups sharing the features being inferred.

AI & Big Data have vastly increased opportunities for profiling. For instance, the likelihood of heart disease of applicants for insurance may be predicted on the basis of their health records, but also on the basis of their habits or social conditions; the creditworthiness of loan applicants may be predicted on the basis of their financial history but also on the basis of their online activity.

A **learned correlation may also concern a person's propensity to respond in certain ways to certain stimuli**. This would enable the transition from prediction to behaviour modification (both legitimate influence and illegal/unethical manipulation).

Example Consider for instance a machine learning system that has learned a model from a training set consisting of previous loan applications and outcomes. The system's *training set consists of personal data.* Correlations embedded in the algorithmic model are not personal data, since they apply to all individuals sharing similar characteristics (*group data*). Assume that this model is then applied to the input data consisting in the description of a new applicant and the default risk attributed to him or her by the model represent personal data.

In the Article 29 Working Party's statement is said that in case of automated inference (profiling), data subjects have the right to access both the input and the (final or intermediate) conclusions automatically inferred from such data. On the contrary, the right to rectification (ask to correct mistaken data) only applies to a limited extend - only if inferred information is 'verifiable'.

Right to reasonable inference: the right that any assessment of decision affecting them is obtained through automated inferences that are reasonable, respecting both ethical and epistemic standards. It has been argues that for an inference to be reasonable it should satisfy the following:

- **Acceptability**: the input data should be normatively acceptable as a basis for inferences concerning individuals (exclusion of prohibited features - sexual orientation).

- **Relevance**: the inferred information should be relevant to the purpose of the decision and normatively acceptable in that connection (ethnicity should not be inferred for the purpose of giving a loan).

- **Reliability:** both input data and the methods to process them should be accurate and statistically reliable.

12.4 Consent

Article 4(11) GDPR - Consent: Consent should be freely given specific, informed and unambiguous, and be expressed through a clear affirmative action:

'consent' of the data subject means any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her.

Recital specifies that consent should be **granular** - it should be given for all the purposes of the processing. Consent plays a key role in the traditional understanding of data protection of data protection, based on the 'notice and consent' model, according to which data protection is aimed at protecting a right to 'informational self-determination'.

Criticism:

- Consent is most often meaningless: not based on real knowledge or the processing at stake, nor on a real opportunity to choose.
- Consent, when targeted on specific purposes, does not include future, often unknown, uses of the data, even when such uses are socially beneficial.

AI and Big Data raise three key issues concerning consent: specificity, granularity, freedom.

Specificity of consent: consent need to be specific, so that it cannot extent beyond what is explicitly indicated. The requirement of specificity is attenuated for scientific research which allows consent to be given not only for specific research projects, but also for areas of scientific research.

Granularity of consent: Consent is presumed not be freely given if it does not allow separate consent to be given to different personal data processing operations. This has two implications: 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.

Freedom of consent: Recital excludes the freedom of consent when the data subject has no genuine or free choice or is unable to refuse or withdraw consent without detriment (patients are told that in order to obtain a medical treatment they must consent that their medical data are used for purposes not needed for the treatment). Consent is not free under situations of clear imbalance. It should not provide a valid legal ground for the processing of personal data in a specific case where there is a clear imbalance between the data subject and the controller (typical when AI and data analytics are applied to personal data).

AI and the Data Protection principles:

- **Article 5(1)(a) GDPR: transparency** - any information addressed to the public or to the data subject need to be concise, easily accessible and easy to understand. A specific aspect of transparency in the context of ML concerns access to data, in particular to the system's training set. Access to data may be needed to identify possible causes of unfairness resulting from inadequate or biased data.
- **Article 5(1)(a) GDPR: informational fairness** - requires that data subjects are not deceived or misled concerning the processing of their data. The data subject should be informed of the existence of profiling and its consequences. Also linked to accountability.
- **Substantive Fairness:** concerns the fairness of the content of an automated inference or decision, which may be summarised by referring to the aforementioned standards of acceptability, relevance and reliability.

- **Article 5(1)(b) GDPR: Purpose limitation -** data should be collected for specified, explicit and legitimate purposes and not further in a manner that is incompatible with those. AI technologies enable the useful reuse of personal data for new purposes, that are different from those for which they were originally collected.

- **Article 29 WP: criteria for compatibility/non-compatibility -** the relevant criteria are the *distance* between the new and original purpose, the *alignment* of the new purpose with the data subjects' expectations and the *safeguards* adopted by the controller to ensure fair processing and prevent undue impacts.

12.5 GDPR, Real Cases and DocX Anonymizer (+)

GDPR in short (2018): for the protection of natural persons with regard to the processing of personal data and for the protection of fundamental rights and freedom. It applies to any organization operating inside the EU or any which offers goods and services to EU. It has 99 Articles and sanctions depending on the gravity of the infringement.

Key ideas:

- **Personal Data:** shall be processed lawfully, in a transparent manner, collected for specified purposes, limited to what it is necessary, accurate and up to date, processed in a secure way to avoid unauthorized or unlawful access.

- **Processing** shall be lawful only if the data subject has given consent, the processing is needed for the performance of the contract with the data subject, is necessary for the performance of a task carried out in a public interest. There are special norms for special categories (ethnicity, political opinion, biometric data).

- Controller shall provide who is he and its contacts, contact details to the data protection officer, the purposes of the processing, the period for which the personal data will be stored.

- **Right to be forgotten - Article 17**: obtain from the controller the erasure of personal data concerning him or her.

- **Right to data portability - Article 20**: to receive the personal data concerning him or her.

- **Article 82**: Any person who has suffered damage as a result of an infringement of the regulation shall receive a **compensation** from the controller or processor.

- **Article 83**: Each individual fine should be effective, proportionate and dissuasive considering: the nature, gravity and duration of the violation; action taken by the data controller to mitigate the damage suffered by data subjects; the degree of responsibility of the controller; the previous violation by the data controller; cooperation with supervisory authority; affected categories of personal data.

Two level of GDPR fines:

- **Lower level:** regard record-keeping, data security and protection
- **Upper level:** regard data protection principles, prohibition of processing sensitive data, denial of data subjects' rights, data transfer to non-EU countries.

Example.

TIM in January 2019 - Between 2017 and 2019 the Italian data protection authority received numerous complaints from several million of individuals for **aggressive marketing**

Violations were inappropriately managing of the call centers hired to make the marketing calls; not updating the list of individuals who had opted out of receiving marketing communications; making the consent to marketing communication a condition for customer to receive discount. In addition TIM apps provided incorrect and not transparent information to users and used invalid method to collect consents.

They were fined for 27million euros to be paid within 30 days and to introduce 20 corrective measures.

A pervasive concept in GDPR is **Data Minimization:** personal data shall be adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed.

Docx Anonymizer is a web service which deals with personal data contained in unstructered docx document. It substitutes sequences of nominative with anonymous IDs. The European Court of Justice decided that, for every preliminary ruling cases presented after the 1st July 2018, natural people's names must be replaced with their initials.

Pseudonymization means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information. The result of pseudonymization are the new file with pseudonymized data and a table with the references between the data subject's personal information and respective identifiers.

Docx Anonymizer helps the data controllers to implement measures which meet the principles of:

- **Privacy by Design:** no personal data is stored in the execution system, after the application has ended.
- **Privacy by Default:** the application does not offer users any option that could mislead them into consenting to the processing of their personal data beyond that carried out by the application.

Tech implementations is a web-app which is responsive, one-page and stateless. It will process docx file since it's the most common format, easy to elaborate, widely used and supported.

Nominative detection is performed with the use a dictionary or by specifying a set of couples surname-name. Detection can be done via regex. It must include regular expression which catch full or partial nominative and its delimiter. There's not offline regex(just-in-time). Should support homonymy and case sensitivity.

Involve the use of the library *docx4j* which allows to transform XML nodes from docx in a Java PlainTexts class and create an EntryPoint.

13 Fairness

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 (overconfidence, loss aversion, anchoring) and the widespread human inability to process statistical data as well as typical human prejudice.

Others have underscored the possibility that algorithmic decisions may be mistaken or discriminatory. Only in rare cases will algorithms engage in explicit unlawful discrimination. Outcomes will be discriminatory due to its disparate impact, since it disproportionately affects certain groups.

Prejudice baked into training set may persist even if the inputs to automated systems do not include forbidden discriminatory features.

In other cases, a training set may be *biased against certain groups*, since the achievement of the outcome being predicted is approximated through a proxy that has a disparate impact on that group.

In other cases, mistakes may pertain to the ML system0s biases embedded in the predictor. Unfairness may also result from taking biased human judgements as predictors (recommendation letters).

Unfairness may derive from a dataset that does not reflect the statistical composition of the population.

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.

In many cases, the best solution consists in integrating human and automated judgments, by enabling the affected individuals to request a human review of an automated decision as well as favouring transparency and developing technologies that enable human experts to analyse and review automated decision-making.

AI deicison making: informational fairness + content fairness of inferences/decisions (avoid prejudice, discrimination).

Imply appropriate mathematical/statistical procedures for profiling, technical and organisational measures to ensure correctness of personal data and secure personal data (potential risks).

13.1 The COMPAS system

It is a risk assessment tool used by American judge to determine the risk of recidivism and appropriate correctional treatment. It is based on statistical algorithms and offenders are classified in 3 categories: high, medium and low risk with a probability score based on a multiple-choice test (137 questions) with static (prior criminal history, education) and dynamic (drug abuse, employment) risk variables.

Loomis case: 2013 - Loomis was charged with driving a stolen vehicle and fleeing from police. He was classified at high risk of recidivism and sentences for 6 years imprisonment and the decision was appealed for violation of due process rights. Since COMPAS functioning is unknown, discriminates both gender and race and its validity can not be verified.

In 2016, the Supreme Court of Wisconsin rejected all defendant's arguments since: statistical algos does not violate the right to individualized decisions, they should be used to enhance judge's evaluation, considering gender is necessary to achieve statistical accuracy. Judges should be informed on the debate concerning COMPAS race discrimination.

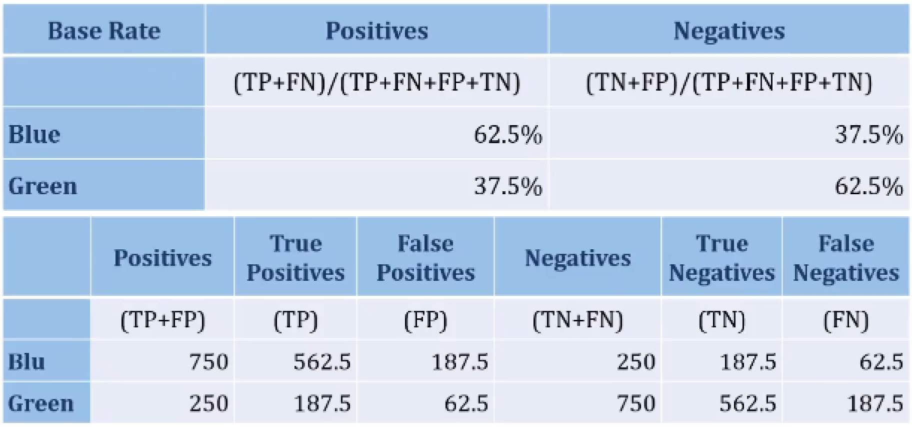
In 2016, **ProPublica** published a **study** to evaluate COMPAS accuracy and fairness. They compared predicted recidivism rates and the rate that actually occurred over 2-year period.

They found a moderate-low predictive accuracy (61%), black defendant were predicted at higher risk - probability of misclassification (45% blacks vs 23% whites). White defendants were often predicted to be less risky than they were - probability of low-risk misclassification (48% whites vs 28% blacks).

Northpoint stated that ProPublica made several errors since the accuracy of COMPAS predictions is higher than human judgments, General Recidivism Risk Scale is equally accurate for blacks and whites. COMPAS is compliant with the principle of fairness. It does not implement racial discrimination.

13.2 SAPMOC case

2000 defendants - 1000 blue and 1000 greens. A single predictor: if previous offences then probably recidivate.Two assumptions: if previous offenders: 75% recidivate, otherwise 25% recidivate. Assume 75% of blues are previous offenders and 25% greens are previous offenders.



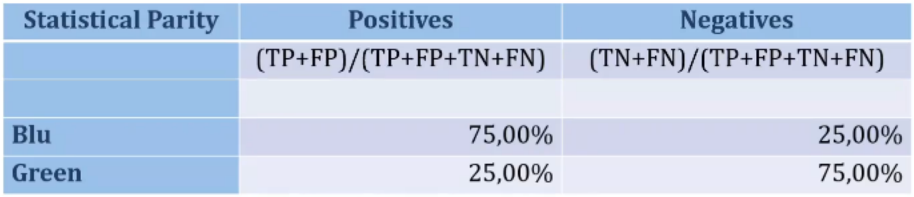
Base Rate	Positives	Negatives
	(TP+FN)/(TP+FN+FP+TN)	(TN+FP)/(TP+FN+FP+TN)
Blue	62.5%	37.5%
Green	37.5%	62.5%

	Positives	True Positives	False Positives	Negatives	True Negatives	False Negatives
	(TP+FP)	(TP)	(FP)	(TN+FN)	(TN)	(FN)
Blu	750	562.5	187.5	250	187.5	62.5
Green	250	187.5	62.5	750	562.5	187.5

SAPMOC accuracy had 75% for both blues and greens which is the same as COMPAS case.

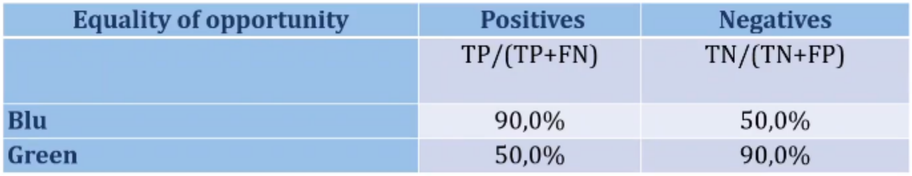
SAPMOC fairness:

- **Statistical parity (NO)**: each group should have an equal proportion of positive and negative predictions.



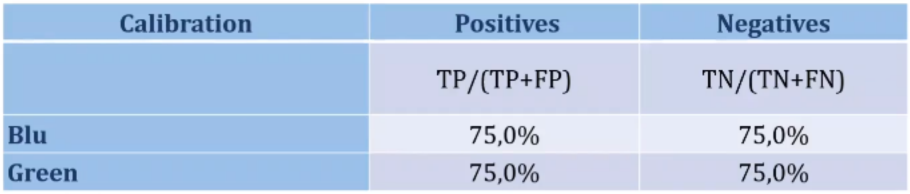
Statistical Parity	Positives	Negatives
	(TP+FP)/(TP+FP+TN+FN)	(TN+FN)/(TP+FP+TN+FN)
Blu	75,00%	25,00%
Green	25,00%	75,00%

- **Equality of opportunity (NO)**: the members of each group, which share the same features, should be treated equally in equal proportion.



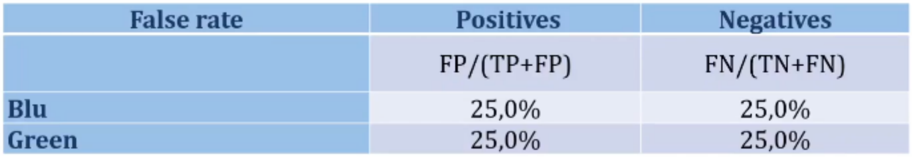
Equality of opportunity	Positives	Negatives
	TP/(TP+FN)	TN/(TN+FP)
Blu	90,0%	50,0%
Green	50,0%	90,0%

- **Calibration (YES):** the proportion of correct predictions should be equal within each group and with regard to each class.



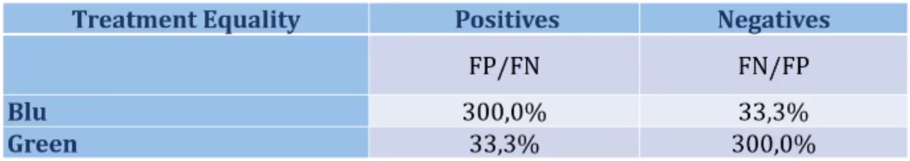
Calibration	Positives	Negatives
	TP/(TP+FP)	TN/(TN+FN)
Blu	75,0%	75,0%
Green	75,0%	75,0%

- **Conditional use error (YES):** the proportion between FP (FN) and the total amount of positive (negatives) predictions should be equal for the 2 groups.



False rate	Positives	Negatives
	FP/(TP+FP)	FN/(TN+FN)
Blu	25,0%	25,0%
Green	25,0%	25,0%

- **Treatment equality (NO):** the ratio between errors in positive and negative predictions should be equal in all groups.



Treatment Equality	Positives	Negatives
	FP/FN	FN/FP
Blu	300,0%	33,3%
Green	33,3%	300,0%

Difference base rate explains the violation of statistical parity, treatment equality and equality of opportunities. Violation of fairness criteria does not necessarily lead to unfairness.

Shall we impose statistical parity: lower accuracy + higher false rate + discrimination against individuals.

Considerations. Unpack the decision: unfairness in prediction (prohibited features, biased data set, biased proxy), unfairness in classification (threshold - affirmative actions), unfairness in decision (right/values optimization).

AI is often perceived as a source of threats and 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.

13.3 Fairness in Automated Decisions (+)

AI bias is an anomaly in the output of machine learning algorithms. These could be due to the prejudiced assumptions made during the algorithm development process or prejudices in the training data.

- **Algorithmic prejudice:** The similarities between protected features and other factors are the source of algorithmic bias. When this happens, removing the protected characteristics from our analysis will not eliminate bias because the correlation may lead to biased decisions based on non-protected factors.

  The Allegheny Family Screening Tool is a model designed to assist humans in deciding whether a child should be removed from their family because of abusive circumstances. Families in the middle and upper classes have a greater ability to "hide" abuse by using private health care providers. African-American and biracial families are referred to Allegheny County more than three times as often as white families.

- **Negative legacy:** Families in the middle and upper classes have a greater ability to "hide" abuse by using private health care providers. African-American and biracial families are referred to Allegheny County more than three times as often as white families.

  In 2019, Facebook was found to be in contravention of the US constitution, by allowing its advertisers to deliberately target adverts according to gender, race and religion, all of which are protected classes under the country's legal system. For example, job adverts for roles in nursing or secretarial work were suggested primarily to women, whereas job ads for janitors and taxi drivers had been shown to a higher number of men, in particular men from minority backgrounds.

- **Underestimation:** Just as data can be biased, it can also be insufficient. Without enough data, machine learning models can fail to converge or provide reliable predictions. This is the problem of underestimation.

  Face recognition algorithms boast high classification accuracy (over 90these outcomes are not universal. A growing body of research exposes divergent error rates across demographic groups, with the poorest accuracy consistently found in subjects who are female, Black, and 18-30 years old.

To minimize bias we need to be aware of the fields in which AI may correct to exacerbate bias; establish processes and practices to test for and mitigate bias in AI systems; engage in fact-based conversations about potential biases in human decisions; fully explore how humans and machines can best work together; invest more in bias research and in diversifying the AI field itself.

Some tools are:

- **What-If:** interactive tool to test, explore and debug ML models.
- **AI fairness 360:** for both detection and elimination of bias.
- **IBM Watson OpenScale:** for enterprises to monitor if their models are acting in a biased manner at runtime.
- **FairlML:** Python toolbox to audit ML predictive models.

Technically, AI can be completely unbiased but not any time soon. The quality of an AI system's input data determines how good it is. You can create an AI system that makes impartial data-driven decisions if you can clean your training dataset of conscious and unconscious assumptions about race, gender, and other ideological concepts. But, as long as there will be biases in our society, models will reflect them.

13.4 Personal Data to avoid discrimination (+)

Evidence suggest that decision making by algorithms may discriminate people, even if the computing process is fair and well intentioned.

We have two contradictory goals. We have to ensure that data-driven decision making is not discriminatory (fairness) and restrict overall collecting and storing of private data to a necessary minimum (general data protection regulation).

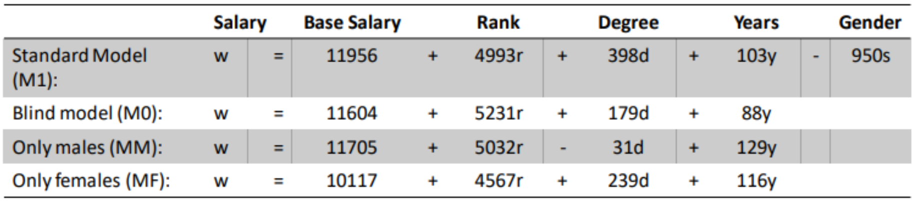
Direct discrimination: based on prohibited characyeristics (race, gender).

Indirect discrimination: based on features that can be proxies for sensitive characteristics. **Masking**: when particular characteristics are found to be correlated, it may be possible to use characteristics as indicators of sensitive characteristics.

An optimal solution can be represented by a direct approach which omits all the unethical constraints inside the model. Omitting those does not make a model free from discrimination.

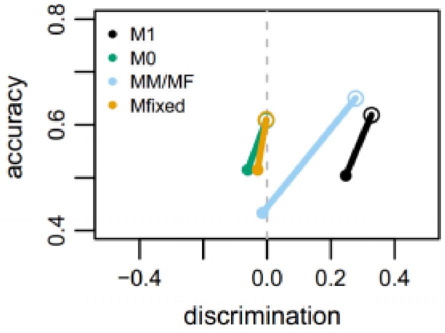
Case study.

We have a dataset of professor salaries with features about rank, degree, gender and years since degree (indicator for experience) and the target is the salary. We may develop different models: M1 a standard model including all variables, M0 a blind model leaving out the sensitive attributes, MM trained only on male data, MF trained only on female data and MFixed a standard model with a constant instead of gender.



	Salary		Base Salary		Rank		Degree		Years		Gender
Standard Model (M1):	w	=	11956	+	4993r	+	398d	+	103y	-	950s
Blind model (M0):	w	=	11604	+	5231r	+	179d	+	88y		
Only males (MM):	w	=	11705	+	5032r	-	31d	+	129y		
Only females (MF):	w	=	10117	+	4567r	+	239d	+	116y		

We can see from results that female model has a lower base salary.



Discrimination is modelled as a **propensity score** estimated using logistic regression to predict gender based on the profile.

We see that the standard and separate models are highly discriminatory, while M0 and Mfixed perform similarly well.

There exists a contradiction between data protection requirements limiting the use of personal data and anti-discrimination law.

Including sensitive personal data in decision-making models may yield less discrimination but is not always allowed under the current data protection law. Solutions:

1. Use anonymous data that is not subjected to personal data protection law

2. exemption to the use of sensitive personal data when used for creating models that explicitly intend to reduce discrimination.

Future steps: Since this method is only applicable to easily explainable models, develop a method that works with complex models.

Pursue the path of a regulatory law to allow for use of sensitive data given the owner's permission, possibly under restrictive conditions to not incur in data leak.

13.5 Fairness through Awareness (+)

The conceptual framework was based on a task-specific metric for determining the degree to which individuals are similar with respect to the classification and on an algorithm for maximizing utility subject to the fairness constraint, that similar individuals are treated similarly.

Fair affirmative action. The aim is guaranteeing **statistic parity**: the demographics of the set of individuals classified are the same of the underlying population.

Fairness limits: datasets are difficult to compare quantitatively, since the notion of fairness can be formalized in different ways, and their scales may differ. In general, decreasing the fairness penalty implies a steady increase of the accuracy, therefore it may be necessary to fix a trade-off between the satisfaction of the fairness constraint and the model accuracy.

Price of Fairness: measure of how the fairness constraint impacts the model accuracy.

- **Statistical Fairness**: have constraints that bind at the level of a category, easy to ensure/check/satisfy. Being an 'aggregated' metric, provides guarantees to the whole group. Measured through Disparate Impact Discrimination index.
- **Individual Fairness**: have constraints that bind at the level of individuals, hard to ensure due to lack of data points, it can be loosen in order to favour members of discriminated groups. Measured in terms of Disparate Treatment DI.

In discrimination-Aware classifiers, we can restrict our domain to a binary sensitive attribute and a binary class without loss of generality. S=b,w where b is the discriminated group and C=+,- where è is the desirable class.

Discrimination of a classifier C with respect to the sensitive attribute b as: $disc_{S=b} = P(C(X) = +|C(S) = w) - P(C(X) = +|X(S) = b$ where X is a random unlabeled data object.

Data Preprocessing techniques:

- **Suppression:** Remove S and the most correlated attributes. Not so effective in fairness scenarios.
- **Massaging:** Swap the labels of some objects in the dataset in order to remove the discrimination from input. Use a ranking model to do so.

- **Reweighting or Sampling**: instead of changing the labels, the tuples in the training set are assigned sample weights. For those who cannot directly work with weights we use sampling, calculate sample sizes for the combinations of S- and C-values, then apply stratified sampling on the groups.

All methods have in common that to some extent accuracy must be traded-off for lowering the discrimination. Those methods, by the way, are practically limited to datasets with categorical sensitive attributes only and single sensitive attribute only. With multiple sensitive attributes becomes more complex. When considering both gender and ethnicity, black females may be more disadvantaged than white females.

Monotonicity Constraints: Such ethical principles can be incorporated into a machine-learned model by adding shape constraints that force the model to respond only positively to relevant inputs helping produce more responsible and trustworthy AI.

- **Unfair penalization:** there may be inputs that a responsible model may reward but should never penalize.

- **Favor the less fortunate:** there may be inputs that help us identify the less fortunate and favor them if there are no other relevant differences.

Non-linear machine learned models can easily overfit noise or learn bias in a way that violates social norms or ethics this problem can be overcomed by training with monotonicity constraints to reflect the desired principle.

Enforcing such constraints addresses deontological (impose rules on how the model can respond to inputs, producing an implicit ethical agent) rather than consequentialist ethics but can also improve or bound (consequentialist) one-sided statistical fairness violations.

Monotonicity constraints are a necessary and useful tool for creating responsible AI however, certainly not sufficient or applicable to all situations.

This will be one of many tools and strategies to achieve responsible AI, therefore consider also using *Tensorflow lattice 2.0* library from now on.

13.6 Fairness in automated decisions (+)

Automated Decision Making Systems ADMS: refer to technical systems that aim to support or replace human decision making.

Definitions for **fairness**:

- Fairness is concerned with actions, processes and consequences that are morally right honorable and equitable. The virtue of fairness establishes moral standards for decisions that affect others.

- Any case where AI/ML systems perform differently for different groups in ways that may be considered undesirable.

- Algorithmic fairness focus on guaranteeing equality of some notion of benefit across different individuals

Group Fairness requires that different groups should have similar outcomes.

Bias means an inaccurate representation of what is true. In a social context, bias is a preference for an outcome about an individual or a group of people. It can be due to:

- Input Data bias: unbalanced data, human bias and proxies.
- Algorithmic bias: depends on the design of the algorithm, data collection and selection.

The deployment of AS in **juridical domain** may bring concerns on systems about automated trial and criminal detection. *In Detroit 2020, a man was arrested and charged for robbery of a boutique store that sells watches, but later the man in the video footage was later confirmed not to be him. So we have a faulty face recognition system.*

How FR systems are faring? It has been observed in multiple instances that many of FR systems currently in commerce have troubles in recognizing women and non-caucasican ethnicities. This systems may suffer of **own-race bias**: people of a different ethnicity may all look the same. From some studies we know that people belonging to different races utilize race-specific features for identifying individuals.

There are face images with facial **makeup** that carry artificial colors, may change skin tone but some approaches try to mitigate the problem with data augmentation suitable for makeup-invariant face recognition.

HireVue is a interview platforms which analyze gestures, pose, lean, eye-contact up the voice tone and content of responses to predict the best candidates. 29% of the score is given by expressions, emotion and personality.

But people with autism, depression, cultures where eye-contact is viewed differently?

It uses also videogame-based assessments, but it is known that younger male have faster reactions.

NLP have demonstrated biases against women and people with disabilities. In speech recognition 35% of the words spoken by blacks were not recognized in (also for women). This because most of the data used comes from TED Talks speeches by white male.

Another case come from **Amazon**'s CV filtering system for software development jobs who penalize the word 'women's' - Amazon abandoned the project.

FairCVtest is an experimental framework which tries to evaluate the capacity to detect protected attributes. It is capable of extracting ethnicity information from the same facial feature embeddings.

In **Financial services**, ADMS may produce price discrimination, targeted advertising and credit scoring.

Price discrimination: shops aim to charge each consumer the maximum price that he is willing to pay. Asians were 1.8 times as likely to be quoted a higher price than non-Asians.

Targeted advertising: selecting a group of people to reach/show ads of products they are willing to buy. Ads about jobs depends on gender. Facebook has been charged with housing discrimination by the US government.

Credit scoring: algorithm discriminate against those who are less willing to share their data online.

Human can increase validity and fairness of data and algorithms. The human can increase clarity and transparency which limits biases in the data introduced by documenting every decision in the process, along with assumptions and hypotesis.

Human can identify misbehaviours undertaken by an autonomous system and take corrective action. Keeping humans in the loop would also provide accountability.

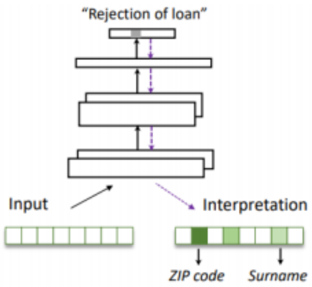
13.7 Deep Learning solutions to fairness (+)

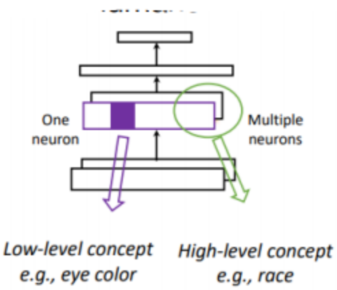
Prediction Outcome Discrimination: DNN models produce unfavourable treatment of people due to the membership of certain demographic groups.

Prediction Quality Disparity: nderrepresentation problem, where data may be less informative or less reliably collected for certain parts of the population.

Interpretability for Addressing Fairness:

- **Local interpretation:** illustrate how the model arrives at a certain prediction for a specific input. The interpretation result is a heat/attribution map of the input.
- **Global interpretation:** provides a global understanding about what knowledge has been captured by a pretrained DNN, and illuminate the learned representations in an intuitie manner to humans.




One neuron
Multiple neurons
Low-level concept e.g., eye color
High-level concept e.g., race

Bias detection:

- Discrimination via Input: local interpretation is employed to generate feature importance vector.
- Discrimination via Representation: global interpretation to identify whether a protected attribute has been captures by the intermediate representation.

A typical deep learning pipeline could be split into three stages: dataset construction, model training and inference. Mitigation methods could be divided into three broad groups: preprocessing, in-processing, post-processing.

Discrimination via input:

- Preprocessing: remove those fairness sensitive features from training data or replace these with alternative values.
- In-processing: using implicit regularization we enforce DNN models to pay more attention to correct features relevant to prediction task, rather than capture spurious correlations between prediction taks and protected attributes.
- Post-processing: takes the model's prediction and protected attribute to calibrate model's prediction during the inference time.

Discrimination via representation:

- Preprocessing: Balanced datasets are not enough, DNNs still could capture gender, race in intermediate representations.
- In-processing: Adversarial Learning to enforce deep representation to maximally predict main task labels, while at the same time minimally predict sensitive/protected attributes.
- Post-processing: suppress the neurons that have captured protected attributes. Global interpretation methods such as CAV are used to locate neurons that are highly related to those - set to zero.

Concept Activation Vector CAV: is the normal to a hyperplane separating examples without a concept and examples with a concept in the model's activation. For the class of interest use directional derivative to quantify conceptual sensitivity. This SC, k, I(x) can quantitatively measure the sensitivity of model predictions with respect to concepts at any model layer.

14 Ethical framework

14.1 AI Ethics at IBM (Francesca Rossi talk)

AI applications: digital assistants (alexa, waze), driving/travel support (tesla, uber, lyft), online recommendations (facebook, amazon, netflix), media and news (google), healthcare (medical image analysis, treatment plan), financial services (credit risk scoring, loan approval, fraud detection), job market (resumee prioritization).

Ethical issues examples: Gender-biased Apple credit card approval process, discrimination in ride-sharing dynamic pricing (uber and lyft), gender-biased recruitment software (amazon), chatbot that exhibited racist speech (microsoft), unethical usage of personal data (facebook and cambridge analytica).

AI ethics: is a multidisciplinary field of study, aims at how to optimize AI's beneficial impact while reducing risks and adverse outcomes, how to design and build AI systems that are aware of the values and principles to be followed in the deployment scenarios and aims to identify, study, and propose technical and nontechnical solutions for ethics issues arising from the pervasive use of AI in life and society.

Main ethical issues:

- **AI needs data**: so data privacy and governance - GDPR
- **AI is often black box:** needs explainability and transparency - AI systems cannot be black boxes, the GDPR limits to decision-making based solely on automated processing and profiling; we have the right to be provided with meaningful information about the logic involved in the decision. It is required by law in Europe. Also we should build different kind of explanations based on the different kind of people who need them (doctors needs more detailed explanations about healthcare AIs)
- **AI can make or recommend decisions:** needs fairness and value alignment
- **AI is based on statistics** and has always a small percentage of error - they are not perfect since they are stochastic
- **AI can profile people and manipulate preferences** - needs human and moral agency. AI can infer our preferences and use them to advertise products that we probably like
- **AI is very pervasive and dynamic**: we may have larger negative impacts for tech misuse and fast transformation of jobs and society. Many jobs will disappear and many others will be created, all jobs will change.
- **Good and bad use of technology**: such as autonomous weapons and mass surveillance. AI can help in achieving the SDGs but the path is very difficult and the pandemic has worsened the situation.

AI is biased since it is trained on data provided by people, and people are biased. There could be bias in the data distribution and in the data labels (ImageNet).

There are many decision points:

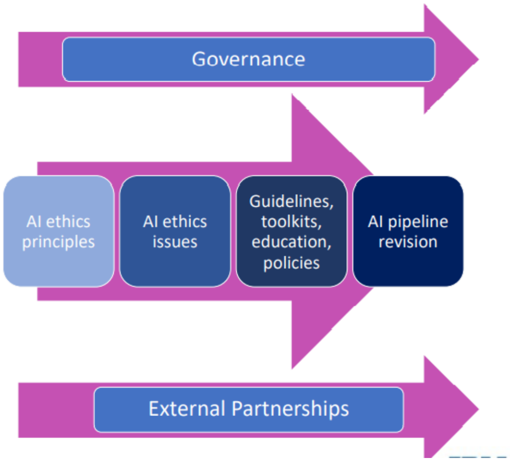
- Individual vs group fairness: similar individuals should receive similar treatments or outcomes or groups defined by protected attributes should receive similar treatments or outcomes.
- Context-dependent definition of fairness
- Acceptable bias threshold
- When to detect bias

IBM Principles of Trust and Transparency (2017): the purpose of AI is to augment human intelligence, data and insights belong to their creator, new technology including AI systems must be transparent and explainable.

What does it mean to trust a decision made by a machine? The four pillars to trust decisions by a machine are: Is it fair or is it going to make discriminatory decisions? Is it possible to understand why it made that decision or is it black box? Is it robust? Is it transparent?

IBM uses technical solutions to detect and mitigate AI bias such as research work, Watson OpenScale, open source libraries (AI fairness 360). Also it offers education and training to developers to adopt new strategies, new frameworks and design thinking sessions.

It also works with neurotechnologies which have a huge potential for healthcare, in reading/writing neurodata, and may raise additional issues around privacy, agency and identity. But works also on quantum computing and so how to responsibly use such a huge computing power.

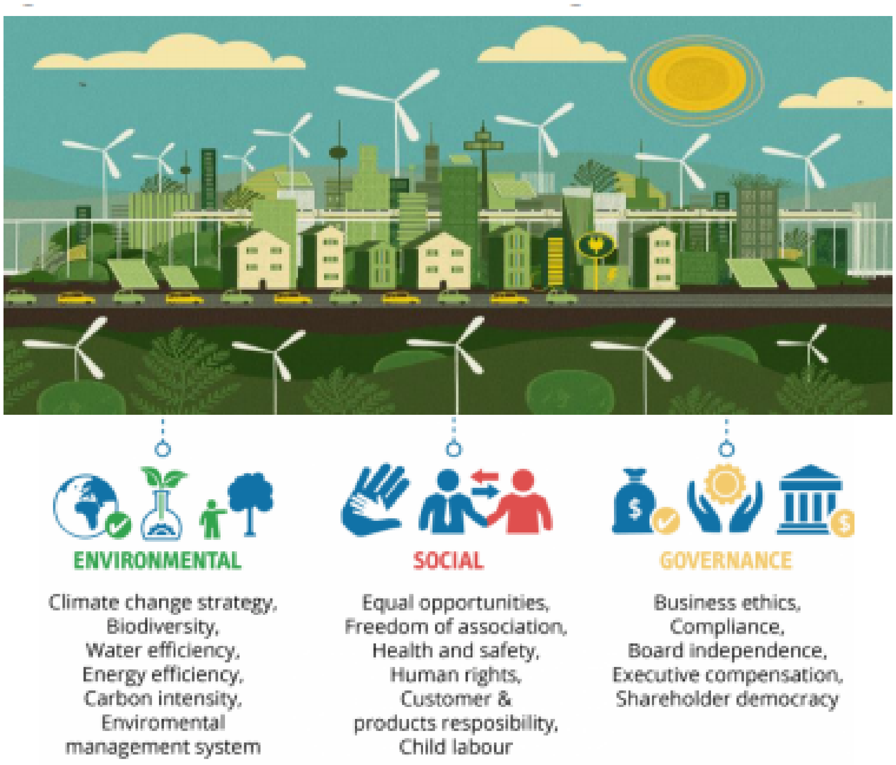


14.2 Based on Sustainable Development Goals (+)

The 2030 Agenda for Sustainable Development, adopted by all United Nations Member States in 2015, provides a shared blueprint for peace and prosperity for people and the planet, now and into the future.

At its heart are the 17 Sustainable Development Goals (SDGs), which are an urgent call for action by all countries - developed and developing - in a global partnership.

The structure of the SDGs is not simple because each of the 17 goals contains targets and each target is evaluated by different indicators. Each goal tipically has 8-12 targets and each target 1/4 indicators.


ENVIRONMENTAL
Climate change strategy, Biodiversity, Water efficiency, Energy efficiency, Carbon intensity, Enviromental management system
SOCIAL
Equal opportunities, Freedom of association, Health and safety, Human rights, Customer & products resposibility, Child labour
GOVERNANCE
Business ethics, Compliance, Board independence, Executive compensation, Shareholder democracy

In order to better study the effects, we can divide the SDGs into three categories, according to the three pillars of sustainable development, namely Society, Economy and Environment.

In this study **AI** is any software technology with at least one of the following capabilities: perception including audio, visual, textual, and tactile (e.g., face recognition), decision-making (e.g., medical diagnosis systems), prediction (e.g., weather forecast), automatic knowledge, etc.

Thanks to the study of about 440 papers, where a consensus-based expert elicitation process was used, we find that AI can **enable the achievement of**

134 goals across all goals, but can also **inhibit 59 goals**. For some goals no papers were found.

SGD Impact Assessment Tool:

1. Gather your forces.
2. Define, refine and draw the line.
3. Sort the SGDs: for each of the 17 macro goals we label them by relevant, not relevant and don't know. If most of the answers are don't know go back to step 2.
4. Assess your impact: we can have direct positive/negative (immediate) or indirect positive/negative impacts.
5. Choose strategy forward.

Test of the tool with some AI projects:

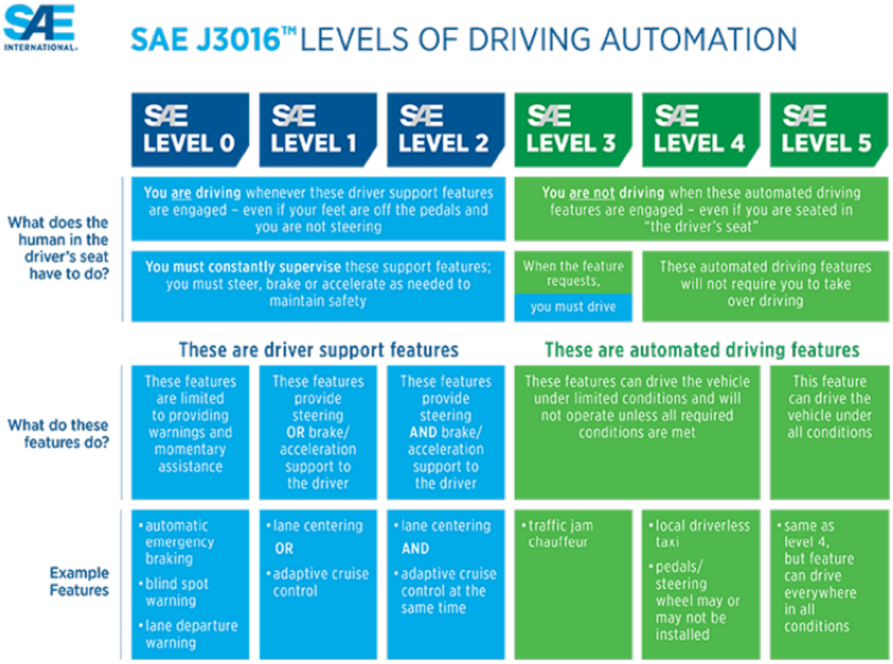
- **Neuron** is an artificial intelligence program that, when paired with an autonomous robotic system called Cortex, can identify and rapidly retrieve recyclable materials on a conveyor belt.

  Impacts directly positively: Sustainable cities community and Responsible consumption production.

- **Neuralink** designed the first neural implant that will let people control a computer or mobile device anywhere. Micron-scale threads are inserted into areas of the brain that control movement. Each thread contains many electrodes and connects them to an implant, the Link.

  Impacts directly positively on Good health and well being, but could impact negatively on Reduce inequalities or even Peace, Justice and Strong institutions.

15 Autonomous Vehicles



SAE INTERNATIONAL

SAE J3016™ LEVELS OF DRIVING AUTOMATION

	SAE LEVEL 0	SAE LEVEL 1	SAE LEVEL 2	SAE LEVEL 3	SAE LEVEL 4	SAE LEVEL 5
What does the human in the driver's seat have to do?	**You are driving** whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			**You are not driving** when these automated driving features are engaged – even if you are seated in "the driver's seat"		
	You must constantly supervise these support features; you must steer, brake or accelerate as needed to maintain safety			When the feature requests, you must drive	These automated driving features will not require you to take over driving	
	These are driver support features			These are automated driving features		
What do these features do?	These features are limited to providing warnings and momentary assistance	These features provide steering **OR** brake/ acceleration support to the driver	These features provide steering **AND** brake/ acceleration support to the driver	These features can drive the vehicle under limited conditions and will not operate unless all required conditions are met		This feature can drive the vehicle under all conditions
Example Features	• automatic emergency braking • blind spot warning • lane departure warning	• lane centering **OR** • adaptive cruise control	• lane centering **AND** • adaptive cruise control at the same time	• traffic jam chauffeur	• local driverless taxi • pedals/ steering wheel may or may not be installed	• same as level 4, but feature can drive everywhere in all conditions

Autonomous vehicles actually look a lot like normal vehicles. There may be radar, gps sensors and cameras on the top of those.

Only level 5 does not include a driver but only passengers.

At the moment we have a kind of soft law for stakeholders, authored by an Independent Expert Group of 14 people mostly academic (philosophy, law, engineering). Established a baseline for future European policy on connected and automated vehicles. 20 recommendations on ethical and social issues.

15.1 Unavoidable collisions

Have been a primary worry in the ethical debate on autonomous driving. How should the system handle morally laden situations - where harm is unavoidable but can be distributed in different ways?

We have the possibility to develop *accident algorithms* which brings a lot of discussions about rights, duties, non-discrimination, sacrifice passengers/bystanders and save more lives possible.

Many issues: Which values/ethical theory to implement? How can we do that? Who gets to decide, how choice should be made, what about personal autonomy and what about the rights of bystanders?

15.2 Privacy & Security

For AV a huge quantity of data must be collected, shared and stored, so we enter the field of privacy issues.

We need to define which data are sensible, so protected.

Also, hacking street signs with stickers could confuse self-driving cars. There's the possibility yo hack self-driving car sensors with a 60$ lidar spoofing device.

15.3 Responsibility Allocation

Who is to be held responsible for harm caused by accidents where AV are involved? Passengers, owners, designers, producers?

Meaningful Human Control Approach: AV must be designed and deployed in a way that assures a satisfying exercise of human moral responsibility + a clear and fair distribution of legal liability → Huge impact on Level 5 Automation.

Has happened that the driver charged in Uber's self-driving cars are charged over fatal crashes but not the company.

15.4 Sustainability

- Environmental Impact: materials are reusable or recyclable, energy consumption, reduce pollution since we reduce traffic.
- Social Impact: more or less traffic not already clear, usable by people affected by disabilities and minorities.
- Economic Impact: will it create new jobs or less jobs, who will be able to afford one of those and is it fair that only rich people can get them.

Value conflicts:

- Safety vs Pleasure
- Personal privacy vs System efficiency
- Moral autonomy vs Human error
- Passenger protection vs Bystanders' rights

Should human driving be outlawed? Yes because it minimize road casualties, and maximize traffic efficiency. No, because of the individual freedom and possible discrimination.

15.5 The Moral Machine Experiment (+)

MME is an online experimental platform designed to gauge social expectations about how AVs should solve moral dilemmas.

What would you want an AV to do if its brakes failed? Keep the lane and hit pedestrians on the road? Swerve and hit pedestrians on the other lane? Hit a barrier with the car?

Analysis was carried out on the following nine dimensions:

- Structural features: staying on course vs swerving, sparing passengers vs pedestrians, sparing more lives vs fewer lives.
- Personal features: sparing humans vs pets, sparing men vs women, sparing young vs elderly, sparing pedestrians who cross legally vs jaywalking, sparing the fit vs the less fit, sparing those with higher social statusr vs lower.

There were created subgroups of Moral Machine users who completed the optional demographic survey on age, education, gender, income and political and religious views.

It was shows that people from different clusters (Western, Eastern, Southern) had different different preferences according to their cultures, perception of law, quality of rules an institutions and demographic reasons.

There were only three main strong preferences: sparing human lives, more lives and young lives.

Weak points: MME concludes that people want AVs to make decisions about who to kill on the basis of personal features, including physical fitness, age, status and gender. The apparent preference for inequality in MME results is driven by the specific "trolley-type" paradigm used by the experimenters.

To prove the ineffectiveness of MME, Bigman and Gray realized 3 different experiments:

1. People were randomly assigned either a forced inequality question (replication of MME in a simplified setting) or an equality-allowed condition ("treat the lives of group A and B equally").

   Result: People overwhelmingly selected the third option when it was available, revealing that they want autonomous vehicles to treat people equally.

2. Similar to the first study, with a modified third option: "AVs should decide who to save and who to kill without considering their personal features"

   Result: Consistent with study 1, people expressed a robust preference for AVs to treat people equally by ignoring personal features.

3. Participants chose which of the two AVs should be allowed on the road: AVs based solely on structural features or both structural and personal features revealed by MME

   Result: 89.9% of participants chose the structural-features-only car, once again expressing a desire for AVs that ignore personal features in ethical dilemmas.

The MME approach allows us to measure the weight of different moral priorities when pitted against each other, rather than considered in isolation, but participants cannot explicitly state that one dimension (for example, age) should not be taken into account.

15.6 The Ethical Knob (+)

Benefits: more safety, less traffic collisions and injuries, easier mobility for children and disabled people.

Obstacles: Disputes on liability, privacy and security concerns, emergence of legal issues and ethical dilemmas.

Scenarios of unavoidable accidents:

A The AV can either stay on course and kill several pedestrians or swerve and kill one passer-by

B The AV can either stay on course and kill several pedestrians or swerve and kill its own passenger.

C The AV can either stay on course and kill one pedestrian or swerve and kill its own passenger.

MES: the same Mandatory Ethics Settings should be implemented in all cars

PES: every driver has the choice to select their Personal Ethics Settings → The Ethical Knob

Ethical pre-programmed AV: An autonomous vehicle could be equipped with pre-programmed approaches to the choice of what lives to sacrifice when losses are inevitable, chosen directly from the AV manufacturers. Despite being a very immediate approach, it establish a paternalistic relationship with the AV user. Furthermore, this solution presents problems concerning the allocation of legal liabilities.

Ethical customised AVs:

An Ethical Knob is a settings mechanism that allows a driver of an AV to delegate the desired moral decision of the previous problem to the vehicle, avoiding paternalization by AV constructors

1. **Altruistic mode:** preference for third parties
2. **Impartial mode:** equal importance given to passengers and third parties
3. **Egoistic mode:** preference for passengers

In scenario (a) passengers' lives are not at stake so the knob's setting doesn't matter. In scenario (b) passengers' lives is at stake and knob setting does matter: save pedestrian; utilitarian approach; always saves own passengers.

Utilitarian approach: adopt the choice that minimizes the sum of the total expected disutilities. We adopt the choice which results in the lower overall disutility.

Rawlsian approach: adopt the choice that minimizes the loss of the most disfavoured individual. By following the rawlsian approach, the AV would choose to swerve and, as a result, risk to kill each of the three passers-by with a probability of 0.6 instead than choosing to kill just one pedestrian with a probability of 0.9.

An Ethical Knob would provide the user with a larger set of choices that can vary over time. Responsibility would be allocated to users, rather than to manufacturers. Consequently, the state-of-necessity defence would work in some cases as it would for drivers in traditional cars. Conclusions The Ethical Knob may improve users' acceptance of AVs, giving them the ability to choose the moral algorithm that reflects their moral attitudes and convictions.

Cybersecurity problems: Phising, DoS, DDos, GPS poisoning, Exfiltration, Communication protocols and rogue updates.

Moral reasoning types:

1. **Moral Altruist:** Type 1 participants can be interpreted as altruistic drivers who emphasize the safety of all involved parties. People in this group tended to make decisions that minimized overall harm done to people by attempting to avoid crashes.

2. **Moral Non-determinist:** Type 2 participants can be interpreted as drivers who make decisions based on context one is situated in. People in this group tended to alter ethical decisions depending on the crash contexts; thus, there were no distinct moral behavioral patterns.

3. **Moral Deontologist:** Type 3 participants can be interpreted as drivers who make decisions based on traffic rules or social norms. People in this group tended to make decisions that followed road traffic rules and moral emotions did not heavily influence the judicial process.

16 Autonomous weapons

16.1 Autonomy

Autonomy:

- a capability that enables a particular action of a system to be automatic or, within programmed boundaries, selfgoverning (US military)
- 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 Bekey)
- an agent's capacity to learn what it can to compensate for partial or incorrect prior knowledge (Russel and 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 and Falcone)

In the concept of autonomy if the standard is too high (comprises all cognitive capacities of humans) then no artificial entity is autonomous. Otherwise, if standard is too low, all algorithms are autonomous.

So autonomy is a **scalable capacity** merging *independence, cognitive skill and teleonomic cognitive architecture.*

1. **Independence**: a technological device, within a system, is independent to the extent that it accomplish on its own, without external interventions a high level task. i.e. land mines or the collision-avoidance system on planes.

   Within a socio-technical system we have an integrated combination of human, technological and organizational components (airplane, manned flying aircraft and civil aviation, autopilot).

2. **Cognitive skills:** an AS engages in high-level cognition (discriminate facts, actions, outcomes) using its own abilities in one or more of the following ways:

   (a) Acquisition and classification of input data (market prices and trends, competition) - Automation with input data, noise reduction and filtering - Sensing pressure light or heat.

   (b) Information analysis to extract further information from input (market analysis, credit rating, voice and natural language understanding) - Automation by computing expected flight trajectories, alert operator of possible risks

   (c) Action selection, construction of plans of actions - Automation by giving suggestions, list of options, take actions

(d) Implementation (scheduling of tasks, perform planned actions and compliance monitoring) - flying according to the established route, monitoring projectile.

i.e. a drone collect information about wind, situation, make forecast about weather and make plan, identify targets, evade threats.

3. **Cognitive-behavioural architecture:**

   (a) **Adaptiveness:** adaptive agents are defined by enclosing boundary that accepts some signals and ignore others, a program inside the boundary for processing and sending signals, and mechanisms for changing the program in response to the agent's accumulating experience.

   An adaptive system can change its patterns of behaviour to better achieve its purposes, in the environment in which it operates - a drone that is able to determine and modify its flight route and possibly even to recognize targets under different environmental conditions.

   (b) **Teleology**: a teleologic system has explicit cognitive states: goals, beliefs (to track aspects of the environment), plans (how to reach goals given beliefs), intentions (as selected plans). These states are differently implemented than corresponding human mental states, simpler but performing the same basic functions.

Multi-layered autonomy: the autonomous behaviour of a system may also emerge from the interaction of lower level non-autonomous or autonomous elements - evolutionary algorithms results from the higher combination of the 'genes' of the most successful algorithms. **Fluid Agency:** agents may be flexibly integrated into higher units of agency through information and decision sharing (fleet of drones or land vehicles).

Human in the loop: autonomy of a device increases as the device is delegated a larger share of the required cognitive tasks - an increased independence of the device, increased interaction/collaboration between the human and the artificial component. Human may remain in the loop while technological devices execute the larger share of the cognitive functions involved in the performance of the task:

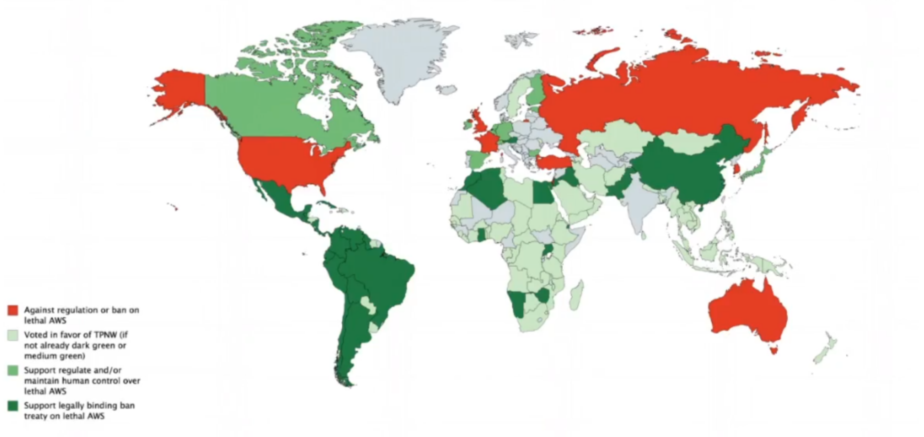
- **Design stance:** describe and anticipate the behaviour of the flight system on the assumption that the system will perform the functions for which it has been designated, as these functions are described in its use-plan

- **Intentional stance:** make the assumption that the system has certain objectives and will make the choices that are best suited to achieve those objectives, given the information that it has.

Cognitive delegation: the delegator chooses to delegate choices instrumental to the execution of a function to the cognitive skills of the delegate system. The delegator doesn't know and thus, doesn't intentionally pre-select what the delegated system will choose to do in future situations.

16.2 Autonomy in Weapon Systems

The US in the 2012 issued a Directive on Autonomy in Weapons Systems. The focus is on **target selection** which involves the determination that an individual target or a specific group of targets is to be engaged.


🟥 Against regulation or ban on lethal AWS
🟩 Voted in favor of TPNW (if not already dark green or medium green)
🟩 Support regulate and/or maintain human control over lethal AWS
🟩 Support legally binding ban treaty on lethal AWS

Autonomous weapons are those once activated, that can select and engage targets without further intervention by a human operator. Those should be used to apply non-lethal and non-kinetic force. Under human supervision, they may engage non-human targets for the defence of manned installations or platforms.

Semi-autonomous weapons systems are intended to only engage individual targets or specific target groups that have been selected by an human operator. May be deployed for any purpose, including the exercise of lethal force against humans, subject only to certification. Drone that send a missile when a person pushed a button.

? (not clear)- Two phases in targeting process in semi-autonomous WS

- Go-onto-target: human delimit the domain of the targets to be selected
- Go-onto-location-in-space: machine selects what particular objects to engage within that domain

Weapons may also rely on a cognitive architecture (the teleological ability to develop plans on how to detect and engage the target, given the available information.

The long-range antiship missile which can't reroute around unexpected threats, search for an enemy fleet, identify one ship it will attack among others in the vicinity, and plan its final approach to defeat antimissile systems - all out of contact with any human decision maker.

Cognition in targeting: targeting process includes all aspects of decision making, which can be automated partially or totally. It involves acquisition of input data from various kind of sensors, information analysis to asses the aspects, features of targets through pattern recognition and computations, decision and action selection for how to engage the target and implementation of the chosen strategy by directing the payload to the destination and possibly monitoring and adjusting trajectories.

Different kind of responsibility:

- **Functional responsibility:** what defect caused the harm - weapon didn't work appropriately
- **Issues of blameworthiness:** did the failura that caused the harm involve a fault
- **Liability:** should somebody pay in tort

16.3 Jus ad bellum & Ius in bello

When is it right to engage a war - only defensive war is legitimate.

Jus ad bellum: when it is legitimate to have military activity against another country

Ius in bello: which concerns how should we behave once we are at war. Purpose it to cause harm to the enemy, to achieve military objectives - when this activity is still legitimate as an action of war.

Four principles of International Humanitarian Law IHL:

- **Necessity**: when harm is caused this must be justified by the purpose that is going to be achieved - there must be a war objective otherwise it would be only unmotivated cruelty (nazists in 2WW). AWS must be allowed to use a limited amount of force in order to avoid unnecessary killings or limit collateral damages but still accomplish the objective of the mission.
- **Distinction**: all military activity should be made against the enemy army, not against civilian population (only by side effect). AWS must be able to distinguish between combatants and not and this is difficult to accomplish because it is very difficult to translate in rules and codes who is a civilian or not.
- **Proportionality**: harm cause to the civilians is proportionate to the military goal being pursued. Harm done must never outnumber the strategic value of the target. AWS don't seem to have a metric to asses the proportionality correctly.
- **Humanity** or *Martens clause*: under any circumstance not explicitly covered by law, use the principles of humanity and public conscience. AWS should have human judgement and interpretation of laws which probably would never be achieved.

Discrimination and *proportionality* could cause an a priori ban on AWS:

- Assuming perfect pattern identification, a **rule-based approach** could guarantee discrimination/distinction.
- Assuming perfect quantification, a **min-max strategy** could yield optimal proportionality.

Precaution and humanity may not imply a ban, but are hard to comply to:

- Assuming perfect communication and real-time planning abilities, it would still be difficult to prevent collateral damage in some cases.
- The humanity principle is so vague and subjective that it may require conscience and human-like intelligence. An alternative would be to limit the autonomy of an AWS and let the humans who have deployed it apply the principle of humanity.

Are autonomous weapons better or worse than humans?

Humans may be cruel, in a situation of rage which can be directed to soldiers or civilians. There is an Humanitarian Law which prohibites to kill a prisoner - if a person surrends he should not be killed.

Some people have argued that AWS would not be subject to those emotions. Also distinction between civilians and non-civilians is difficult to achieve on AWS which is a cons.

It would be just another harm race - where more and more sophisticated AWS would be developed to prevent that the enemy has more advanced AWS → more deadly weapons.

There is a prohibition on chemical weapons/lethal gases. Also, the development of generic AI techniques can be misused for the development of more sophisticated AWS, so we can't stop the development of those in some way.

We should aim at a *symbiotic partnership* between humans and machines, which, not only will perform intellectual operations much more effectively than man alone can perform them but will also perform such operations better than machines alone.

Liability gap: impossibility of attributing moral responsibilities (blameworthiness) and legal liabilities to anyone for certain harms caused by the systems' autonomous operation.

The impossibility of attributing moral responsibilities (blameworthiness) and legal liabilities to anyone for certain harms caused by the systems' autonomous operation.

This liability gap is the main ethical concern that distinguishes conventional weapons from Autonomous Weapon Systems.

Functional responsibility: any component and subcomponent of a sociotechnical system may fail to exercise its function as expected. As a consequence, the system as a whole may fail, with harmful consequences.

Blameworthiness: failure that caused the harm involves a fault, namely a substandard behaviour in a moral agent. It is hardly applicable to automated devices, even when they are autonomous, but it may concern their designers and integrators.

Solutions to liability gap: Schulzke (2013) holds that existing mechanisms for responsibility attribution can be extended to autonomous weapons. The problem of determining responsibility for autonomous robots can be solved by addressing it within the context of the military chain of command.

Arkin et Al. (2019) propose possible actions to address concerns about AWS. In particular, to define and universalize guiding principles for human involvement in the use of force: Militaries must invest in training, education, doctrine, policies, system design, and human-machine interfaces to ensure that humans remain responsible for attacks; Humans responsible for initiating an attack must have sufficient understanding of the weapons, the targets, the environment and the context for use to determine whether that particular attack is lawful.

If something goes wrong (a drone hits a terrorist and an unacceptable amount of collateral damage is done then nobody other than the machine is responsible).

AWS can be used for Law Enforcement to disable mines, block passages or to track people being hostages of criminals and attack exclusively them, without harming hostages.

16.4 Assessing responsibilities (+)

Accountability of war crimes is important for three main reasons:

1. to comply with Jus in Bello principles
2. to avoid the same situation in the future
3. for peace-making after the conflict

State responsibility: states are held responsible of any unlawful action of their army. However, the current IHLs mention liabilities of physical people and don't make any reference to AWS liabilities. Blaming groups can only be reflected into economic sanctions.

International Criminal Laws are concerned with the liabilities of physical entities w.r.t. serious violations of IHLs, a LAWS however is not a person punishable on its own. A complex system is made by many programmers, pinpointing who made the 'defective' part is impossible. If the company made clear the existence of limitations or bugs, the responsibility is assumed by the users.

Moreover, a company can be accounted only for commercial malfunction, not war crimes, unless malicious intent can be demonstrated.

While accountability of a single event is difficult to asses, a LAWS repeating a war crime multiple times makes a **commander** liable of not disabling it to prevent known criminal behaviour.

Accounting for responsibility, means that the culprit must be punished in a way that discourages repeating the same crime in the future. This would make LAWS moral entities similar to humans and would have the same ethical status of a human soldier.

16.5 Cyber Warfare (+)

Cyber warfare is the use of digital attacks against the enemy. This include espionage, sabotage, propaganda (Cambridge Analytica) and economic disruption (ransomwares).

IHL cover the use of cyber attacks only during an armed conflict, however most of the cyber warfare is conducted during peace time. This also bypasses the fact that lack of discrimination should make cyber weapons inherently forbidden by IHL.

Cyber weapons are fully autonomous: they can spread on their own over the internet, they can select a target based on installed software, they can be completely invisible.

A sabotaging weapon such as Stuxnet could disable hospital facilities or cause a missile silo to fire. An information warfare based weapon could spread fake news causing people to kill each other. So, they can be lethal.

Ethics of cyber weapons:

- Attack didn't lead to death (unlike bombing of Iraqi nuclear plants in 1981), so as Stuxnet was more proportional.
- Stuxnet was used without a war declaration
- Stuxnet deliberately contained many safety mechanisms to avoid spreading
- but they didn't work as expected and the virus spread anyway
- Collateral damage is virtually non-existent, because physical harm from cyber weapons must be intentional
- They can escape control more easily
- Cyber weapons effectiveness relies on secrecy, so no explainability is possible
- In general, cyber weapons avoid deaths but depends on the case

In conclusion, cyber weapons are controversial but better. In general they are invisible and responsibilities cannot be assessed. They can escape control easily and don't cause death. They are deployed at peace time and no law regulates them.

16.6 Human-Machine relationship (+)

- **Human-in-the-loop:** also known as *semi-autonomous* are machines that perform a function autonomously for some period of time, then stop and wait for human input. These systems use autonomy to engage individual targets or group of targets that a human has decided are to be engaged (guided munitions, radars) - they do not raise new issues since they are more precise.

- **Human-on-the-loop:** also known as *human-supervised* are machines that can perform a function entirely on their own but have a human in a monitoring role, with the ability to intervene and halt the system operations. Mostly used in defensive situations in which reaction time would make physically impossible for humans to remain 'in the loop'.
- **Human-out-of-the-loop:** also known as *fully autonomous* are machines that can perform a function entirely on their own with humans unable to intervene. There are very limited number of those (loitering munitions and encapsulated torpedo mines). Human doesn't know which target is going to be hit, he only knows the geographic area and the classes that could be hit.

In warfare situation, humans have difficulties to adhere laws because they are put into stressful situations in which fear and instinct of self-preservation prevails over dutifulness and wisdom. Weapons can be used in a self-sacrificing manner.

16.7 AWS violate Human Rights (+)

AWS may violate various human rights:

- **Right of the bodily integrity:** it includes the right of life, right to security, right against cruel, inhuman and degrading treatment - these are violated with the loss of life and accountability is not provided.
- **Right of dignity:** this right is violated because those injured or killed by AWS, their life is affected by an entity unable to do moral judgement therefore unable to be moral, hence their dignity is violated by an agent that does not have dignity.
- **Principle of necessity:** since machine will use lethal or less-lethal force, they can't apply this principle since they do not have any other instruments to achieve it.
- **Principle of proportionality:** violated since AWS are going to applu the indication learned and imposed during development and are unable to react to unusual situations.
- **Principle of distinction:** violated due to inability to distinguish offenders from non-offenders. They may be unable to distinguish a surrender or act accordingly to the personal information of the subject.
- **Principle of accountability:** since accountability in unlikely to be respected, right to life, dignity, remedy, justice and many others will be violated.

16.8 Our presentation (+)

Target recognition methods:

- **Shape detection:** makes it possible to recognize objects in an uncluttered environment - medium to high cluttered environments introduce an unacceptably high false alarm rate

- **Thermal imaging:** detects heat radiating from an object and shows its movement - could not be used to distinguish between a combatant and civilians.

- **Radiation detection:** used by loitering munition to detect radar signal and determine if they are friendly - the radar should be part of an anti-aircraft installation in order to determine the legitimacy of a target.

- **Acoustic direction finding:** calculates the location of the sound by using differences between the times that sound reaches two or more separated microphones - other acoustic effects may be detected and responded to

Even an improved ability to recognize targets does not enable machines to assess whether a target is legitimate and whether the attack as a whole is permissible.

Thus, a more appropriate use of these methods would be to assist human supervisory control to achieve more precise and accurate targeting by a human exerting the power of deliberative reasoning and judgement.

Because humans sometimes fail at some tasks, it does not mean that machines can do them any better. It can simply mean that humans are being asked to perform in a mode of operation that is not well suited to human psychology. This needs to be part of the equation of ensuring efficient and meaningful human supervisory control of weapons.

If we get the balancing act right we could have more precision and accurate targeting with less collateral damage and better predictable compliance with International Humanitarian Law. But getting it wrong could result in considerable humanitarian problems.

- **Deliberative reasoning:** *Your conscious self – the thinking you.* Kahneman D. Goes along with the automatic processes unless there is something surprising or irregular and/or we are operating in novel circumstances or performing tasks that require vigilance and/or deliberation. Requires attention and free memory space. Stress or distractions could incapacitate it

- **Automatic reasoning:** Does not require active control or attention. Can be trained through repetition and practice on routine tasks. Used anytime for routine decisions that have to be made rapidly for predictable events. Works well in environment that contains useful cues that, via practice, have been (over) rehearsed.

Automatic reasoning is problematic because:

1. **It neglets ambiguity and suppresses doubt**: they are guided by experience. It does not search for alternative interpretations and does not examine uncertainty.

2. **It infers and invents causes and intentions:** It is adept at finding a coherent causal story to link together fragments of available information.

   Assimilation bias: a person who is presented with new information that contradicts a preexisting mental model, assimilates the new information to fit into that mental model.

3. **It is biased to believe and confirm:** It favors the uncritical acceptance of suggestions and maintains a strong bias.

   - **Automation bias:** a computer-generated solution is accepted as correct, human decision maker disregards or does not search for contradictory information.

     Error of omission: humans fail to notice problems because the automation does not alert them.

     Error of commission: humans erroneously follow automated directives or recommendations.

   - **Confirmation bias:** when seek out information to confirm a prior belief and discount information that does not support this belief

4. **It focuses on existing evidence and ignores absence evidence:** It builds a coherent explanatory history without considering any evidence or contextual information that might be missing - WYSIATI: What You See Is All There Is - It makes us confident to accept information as true whether it is or not.

The unpredictable and unanticipated circumstances in a dynamically changing environment play to the weakness of automatic reasoning.

It is vitally important that deliberative reasoning is enabled in the design of supervisory control for weapons systems. Although this is also subject to error and flaws, it does as good a job as can be done with uncertainty and doubt.

If a supervisory weapons operator is distracted by another task or if they are stressed, their attentional capacity may be low. They would not be able to use their deliberative reasoning and could simply catch the downsides of automatic reasoning if there were problems or irregularities.

Levels of human supervisory control:

1. **Human deliberates about a target before initiating any attack:** Acceptable if requirements are met: full contextual and situational awareness of the target area; be able to perceive and react to any change or unanticipated situations; active cognitive participation; sufficient time for deliberation on the nature of the target, its significance in terms of the necessity and appropriateness of attack; means for the rapid suspension or abortion of the attack.

2. **Program provides a list of targets and human chooses which to attack:** Acceptable - if requirements of deliberative human control met and no ordered list is provided, otherwise there would be a tendency to accept the top ranked target unless sufficient time and attentional space is given for deliberative reasoning.

3. **Program selects target and human must approve before attack:** Unacceptable - Creates an automation bias in which human operators come to accept computer generated solutions as correct and disregard or don't search for contradictory information.

4. **Program selects target and human has restricted time to veto:** Unacceptable since it does not promote target identification, short time to veto would reinforce automation bias and leave no room for doubt or deliberation and the time pressure will result in operators falling foul of the four downsides of automatic reasoning.

5. **Program selects target and initiates attack without human involvement:** Unacceptable since there is no human involvement in the target selection and attack and no compliance with international law.

SARMO weapons (Sense and React to Military Objects)

Weapons systems in use that operate automatically once activated: they intercept high-speed inanimate objects automatically. They complete their detection, evaluation and response process within a matter of seconds. Difficult for human operators to exercise meaningful supervisory control once they have been activated other than deciding when to switch them off.

Features: small set of defined actions repeatedly and independently of external influence or control; highly structured and predictable environments; switched on after detection of a specific threat; unable to dynamically initiate a new targeting goal or change mode of operation; constant vigilant human evaluation and monitoring for rapid shutdown; predictable output and behaviour; only used defensively against direct attacks by military objects.

From the perspective of the human supervisory control framework, the human decision of when to use the weapon is the key to the legality of SARMO weapons systems.

Conclusions: Fully autonomous weapons could not be used in a way that could be guaranteed to predictably comply with International Law.

There is general agreement on the inadequacy of Automatic Target Recognition. Both humans and computer systems have their strengths and weaknesses: exploit the strengths of both! Do not ask humans to perform in a mode not well suited to human psychology. Do not use computerized weapons that are not meaningfully controlled by human operators.

Develop a principle of human control founded on human reasoning processes to provide clear guidelines for state weapons reviews.

16.9 Arguments against AWS (+)

1. **Limitations of technology:** can we program social awareness? Willt here be room for surrendering? What distinguishes combatants from civilians?

2. **Deontological arguments:** is it ethically defensible to give a machine the right to take a human life?

   - The **accountability gap**: who is responsible for the end of a human life? Operator vs supplier vs high command.
   - The **right to a dignified life:** AWS threaten human dignity since they have to understand the value of life, reason in justice, morality or law and posses the concept of kantian dignity and equality. But do wars in general preserve dignity? Is is ok to have death by an algorithm. What is dignity, human rights?

3. **Consequentialist reasoning:** how will the use of AI change the future international warfare? Is it morally right to allow AWS based on its consequences?

17 Ethics of filtering

Filtering: The classification of user-generated content for the purpose of determining its accessibility: excluding it from all users, making it inaccessible to certain categories of them, upgrading or downgrading its priority.

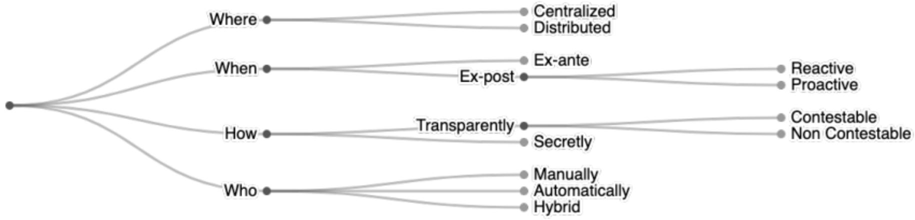
Pros: Moderation in online communities. Blocking unlawful content. Harm prevention. Recommender systems that guide us through the vastness of the web.

Cons: Filters are not perfect. Not transparent. Privacy intrusion. Behaviour manipulation. Filter Bubbles.

17.1 Taxonomy

Moderation: is the active governance of platforms meant to ensure interactions among the users that are productive, pro-social and lawful.

Taxonomy



WHERE:

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

WHEN:

1. **Ex-ante filtering**, which is applied before the content is made available on the platform.
2. **Ex-post filtering**, which is applied to the content that is already accessible to the platform's users.

   - ***Reactive filtering***, which takes place after the issue with an item has been signaled by users or third parties
   - ***Proactive filtering***, which takes place upon initiative of the moderation system, which therefore has the task of identifying

HOW:

1. **Transparent filtering**, which provides information on the exclusion of items from the platform.

   - ***Contestable filtering***. The platform provides uploaders with ways to contest the outcome of the filtering, and to obtain a new decision on the matter
   - ***Non-contestable filtering***. No remedy is available to the uploaders.

2. **Secret filtering**, which does not provide any information about the operation.

WHO:

1. **Manual filtering**, which is performed by humans.
2. **Automated filtering**, which is performed by algorithmic tools.
3. **Hybrid filtering**, which is performed by a combination of humans and automated tools.

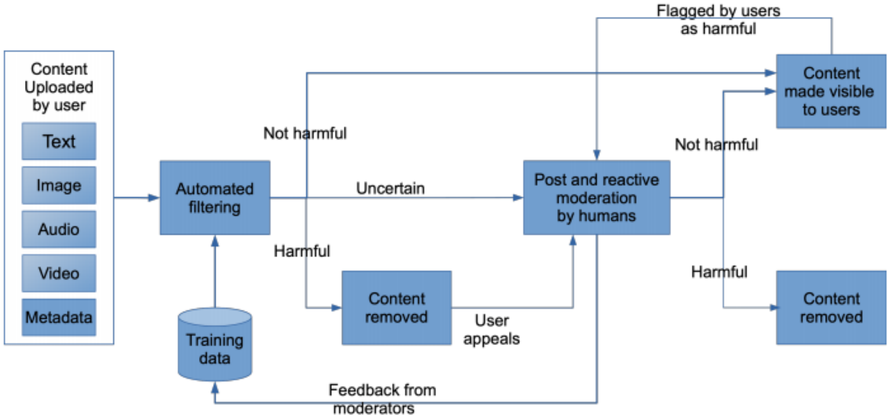
17.2 How it works

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.



All the media is captured by the use of any platform to publish content. There is a first level of automated filtering based on ML techniques. Then it is classified as harmful (content removed), uncertain (post and reactive moderation of humans) or not harmful- which is really prone to error since based on ML approaches which are probabilistic/statistical methods.

Some epic fails involve Facebook who, for example banned Neptune statue photo for being explicitly sexual or Youtube who removed videos showing atrocities in Syria.

17.3 Santa Clara Principles

1. Companies should publish the numbers of posts removed and accounts permanently and accounts permanently or temporarily suspended due to violations of their content guidelines.
2. Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.
3. Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.

17.4 Filter Bubble issue (+)

Filter bubble: State of intellectual isolation that can result from personalized searches when a website algorithm selectively guesses what information a user would like to see based on information about the user, such as location, past click-behavior and search history.

If the hypothesis of filter bubbles is true, we have that people on the internet have the illusion to see everything, to interact with everyone, but in reality they are isolated inside a bubble.

Echo chamber: Metaphorical description of a situation in which beliefs are amplified or reinforced by communication and repetition inside a closed system. An environment in which a person encounters only beliefs and opinion that matches one's own - causes are selective exposure and recommender systems.

When echo chambers are created through recommender systems, they are called filter bubbles.

Explicit personalization: The personalization is tailored to the user based on criteria that the user has explicitly set.

Implicit personalization: The personalization is tailored to the user based on information collected from observing the user's online behaviour.

Cons of filter bubbles

- **Polarization:** Filtering of different points of view leads to an increase in social polarization between communities.
- **Information blindness:** Limited exposure to new informations might narrow our outlook, making us more vulnerable to manipulation and propaganda.
- **Confirmation bias:** Being involved only with information that supports one's values might enhance already strong beliefs and lead to estremism.

Personalized filtering might push people towards a detrimental spiral of attitudinal reinforcement which may cause the rise of extreme viewpoints. The lack of exposure to diverse content, especially in the political domain, may reduce the subject's capability in developing a critical opinion about various matters.

17.4.1 Google News experiment

An experiment on Google News has been made and in the result of this, no relevant differences have been found in the Google News results of the 5 agents. This experiment gave no evidences of the filter bubble phenomenon.

The experiment lasted 1 week and each day was searched for 5 agent-specific terms on the Google search engine and click on the first three results. Then it was expressed approval towards specific content through Facebook's Like button. All this to test the implicit personalization effects on Google News.

17.4.2 TikTok experiment

A similar experiment was made on TikTok. We have analyzed what are the parameters on which the TikTok algorithm is based to customize the contents, in order to have finite and measurable quantities on which to base our experiment and therefore interpret our results. These parameters are for example the likes posted by the user, comments and so on. But also user device and account settings like language preference, country setting and device type.

After a week of use, we can see the effect of the personalization.

Videos of the female account are very different from that of the male account (Gender profiling) and there is a lot of content customized according to the interest of the agent.

This experiment gave no strong evidences of the filter bubble phenomenon.

As we said in the other experiment, the main reasons that led to this conclusion could be two: the filter bubble phenomenon is not so critical, as argued also in recent work. Our experiment failed in deepening enough the implicit personalization.

17.4.3 Study on social networks (Facebook, Twitter, Reddit, Gab)

Platforms organized around social networks and news feed algorithms, such as Facebook and Twitter, can favor the emergence of echo chambers.

Gab didn't present signs of echo chambers probably because they were observing the dynamic inside a single echo chamber.

A clearcut distinction emerges between social media having a feed algorithm tweakable by the users (Reddit) and social media that don't provide such an option (Facebook and Twitter).

The feed algorithms may have a role in the formation of echo chamber effects corroborating the filter bubble hypothesis.

This may lead to polarization and ideological isolation, political polarization, influencing people and easing fake news spread.

17.4.4 Filtering in actuality

All the public informations about filter algorithms implementation are outdated, and the algorithms are in constant evolution: PageRank (Google) and EdgeRank (Facebook) are not working as it once was.

Some counter measures were adopted from medias and search engines to prevent in part the bubble: Google search engine now acts differently then how Pariser described.

The actual major problem shared from the researchers on the filter algorithms is the opaqueness and explainability.

Despite the global behaviour of the filter is predictable, the results of these algorithms are indeed irreproducible and often are neither well understood from their developers

Anti-Profiling methods: signed-out searches, third-parties cookies removal, incognito mode, non-profiling search engines, virtual profiling deception, use of VPN.

Leaving the bubble: open dialog medias, shared search with browser plugins.

An ethical filter what should contain?

Transparency, user customization, content diversification, favor higher quality content visualization over user time permanence or incomes, fake news or ad-hoc content detection.

17.5 Filtering techniques and censorship (+)

- **Text**

  1. **Blacklist**: dataset of unwanted textual content and simply match the words in the text. Used for violation of copyright
  2. **NLP**: a more powerful technique as it analyzes not only the matching but also the semantic, the meaning. Used for illegal hate speech.

- **Image**

  1. **Hash Matching**: construction of hash and fingerprints databases of undesired images.
  2. **Object Recognition**: detect instances of semantic objects of a certain class (such as humans, buildings, or cars) in digital images and videos
  3. **Semantic Segmentation**: process of partitioning a digital image into multiple segments (sets of pixels, also known as image objects) to simplify the representation of an image into something that is more meaningful and easier to analyze
  4. **VQA:** Visual Question answering

- **Audio**

1. **Hash Matching:** construction of hash and fingerprints databases of undesired audio
2. **Speech-to-Text and Speech-Recognition**: convert audio into textual data and use Deep Learning techniques to understand it
3. **Audio Classification**: ML / DL techniques that classify an audio file in order to decide whether it has to be removed or not

- **Video**

    1. Images and audio techniques are combined in order to process and filter videos

17.6 Enemies of the Internet (+)

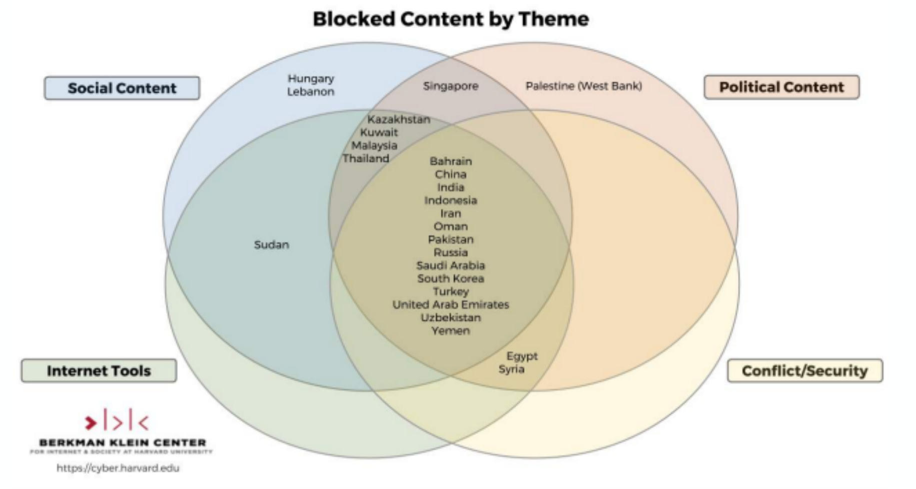
- **Myanmar**: on 4th February 2021: Facebook, Facebook Messenger and WhatsApp were blocked, followed the day after by Twitter and Instagram. While on 15th February the military has imposed nationwide internet blockades from 1 AM to 9 AM and mobile internet was shut down
- **North Korea**: Internet accessible only by senior government officials. Ordinary citizens are restricted to local intranet. VPN and other ”not traceable” technologies are banned.
- **Cuba:** Before free access to the Internet, there were 80,000 illegal telecommunications systems in the homes of citizens. Now the Cuba telecommunication system is under Government monopoly, with unfair costs for the common people: ”economic” internet restrictions. Only in 2018 the state granted access to the 3G network. And in 2013 the government granted the web access through state-owned internet cafes.
- **Belarus:** In December 2018, a law was approved that obliges citizens to enter their full name and mobile number on the network every time they comment on a news item
- **Saudi Arabia:** Heavy filtering and censorship on web content

17.7 Censorship (+)

Censorship is the suppression or prohibition of any parts of books, films, news, etc. that are considered offensive, politically unacceptable or a threat to security.

Censoring methods:

- Packet filtering
- Internet Service Provider (ISP) Filters
- Autonomous System Number (ASN) Blocking
- Network Disconnection and Connection Reset



Reasons behind censorship: Protecting national security. Blocking culturally sensitive material. Maintaining a high level of productivity. Cultural and religious concerns. Regional and internal political conflicts. Benefits of the economy.

Facebook, Instagram, Youtube and all big companies are using machine learning and deep learning for detecting sensitive contents on their websites. Definition of sensitive contents are different in non-democratic countries. Religious, Sexual and Political content filtering are nowadays automated by AI.

17.8 Ethical Recommendation Systems (+)

General idea: filtering based on cultural and ethical preferences. "Ethical database", based on a region's cultural norms. **User-centred design**, with adjustable tools to explicitly control personal data usage, biases filtering, content censorship and online experiments. **Multi stakeholder design**, which focuses on other involved agents as well.

User-centred approach:

- Pros: Minimizes the negative impact of RS on user's utility and user's rights.
- Cons: May be insufficient to protect user's privacy, may result in inefficiency. They shift the responsibility for the protection of rights and utility to the users.

Multi stakeholder approach: has better evaluation metrics of RS, better understanding of the objectives to maximize, consequentialistic approach, need to take into consideration the time frame.

Privacy risk occur in at least four stages:

- When data are collected or shared without the user's explicit consent

- Once data sets are stored, they may be leaked or de-anonymized
- Indirect inferences that can be drawn from the data
- **Collaborative filtering**: the system can construct a model of the user based on date on other users' interactions

Approaches to solve privacy can be *architectural* (with separate and decentralized databases), *algorithmic* (encryption) and *policy* (explicit guidelines and sanctions).

Three types of fairness in RS:

1. **User/consumer C-fairness:** Take into account the disparate impact of recommendation on protected classes of recommendation consumers
2. **Provider P-fairness:** Fairness needs to be preserved for the providers only
3. **Both CP-fairness:** Fairness is considered for both consumers and suppliers

18 CLAUDETTE System

How to empower consumers? We need to protect them against unwanted monitoring (GDPR). Support them in detecting unfair use of AI and help them by controlling commercial practice fairness.

"An opposing exercise of power is the principal solvent of economic power, the basic defence against its exercise in economic affairs"

In the AI era an effective countervailing power needs to be supported by AI.

Claudette is a ML system to automatically detect potentially unfair clauses in Terms of Services and Privacy Policies.

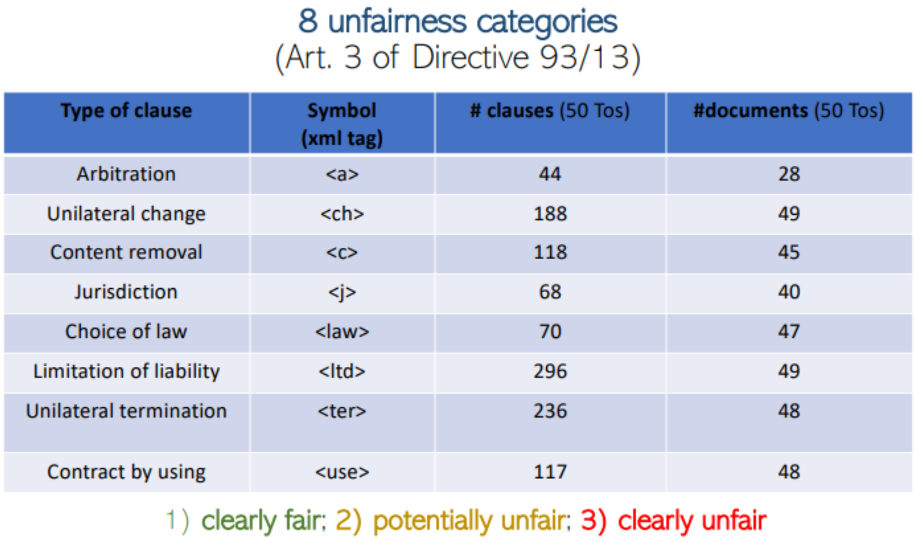
Usually consumers agree but don't read. Business keeps using unlawful clauses even if there are regulations. NGOs have competence to control but lack resources.

Training set (ToS): started from 50 Terms of Service manually annotated, now 100. Over 7090 sentences, there are 787 (11.1%) sentences labeled as unfair clauses.

Directive 93/13 art 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.

The idea is that there are some types of clauses that traders are prohibited from using in the contracts.

- Arbitration
- Unilateral change: clauses stating the service provider has the right to change/modify the content
- Content removal: possibility for provider to remove some content
- Jurisdiction: court where you go in case of issues
- Choice of law: law that is applicable in case
- Limitation of liability
- Unilateral termination
- Contract by using



8 unfairness categories
(Art. 3 of Directive 93/13)

Type of clause	**Symbol (xml tag)**	**# clauses** (50 Tos)	**#documents** (50 Tos)
Arbitration	<a>	44	28
Unilateral change	<ch>	188	49
Content removal	<c>	118	45
Jurisdiction	<j>	68	40
Choice of law	<law>	70	47
Limitation of liability	<ltd>	296	49
Unilateral termination	<ter>	236	48
Contract by using	<use>	117	48

1) clearly fair; 2) potentially unfair; 3) clearly unfair

Consent by using clause: If a clause states that the consumer is bound by the terms of service simply by visiting the website or by downloading the app, or by using the service: **potentially unfair**. (AirBNB and Facebook)

Jurisdiction clause: If giving consumers a right to bring disputes in their place of residence: **clearly fair.** If stating that any judicial proceeding takes a residence away (i.e. in a different city, different country): **clearly unfair.** (Dropbox resolve claims only in San Francisco country - ¡j3¿ xml tag)

Limitation of Liability: If stating that the provider may be liable: **clearly fair**. 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": **potentially unfair**. If stating that the provider will never be liable for physical injuries (health/life)// gross negligence// intentional damage: **clearly unfair**.

Fair - World of Warcraft: Blizzard is liable in case of any injury to life, limb or health. Potentially unfair - 9gag: discontinuance or lack of availability may cause damage to the consumer. Clearly unfair - Rovio: not liable of anything.

From an ML point of view, the problem is modelled as a:

- **Detection task:** does a sentence contain a potentially unfair clause
- **Sentence classification task:** what is the category the unfair clause belong to

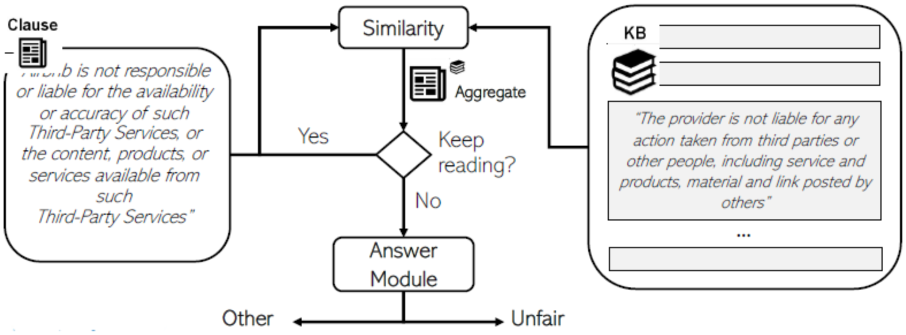
Approaches: Bag of Words, Tree kernels (structure of sentences by describing the grammatical relations between sentence through a tree), CNN, SVM.

Leave-one-out procedure used: each document in turn is used as test set, leaving the remaining documents for training and validation set for model selection.

The *best model* found is an ensemble of multiple classifiers with an average of 80% of correctly classified potentially unfair clauses. It has been developed an online automated detector of potentially unfair clauses with the respective category.

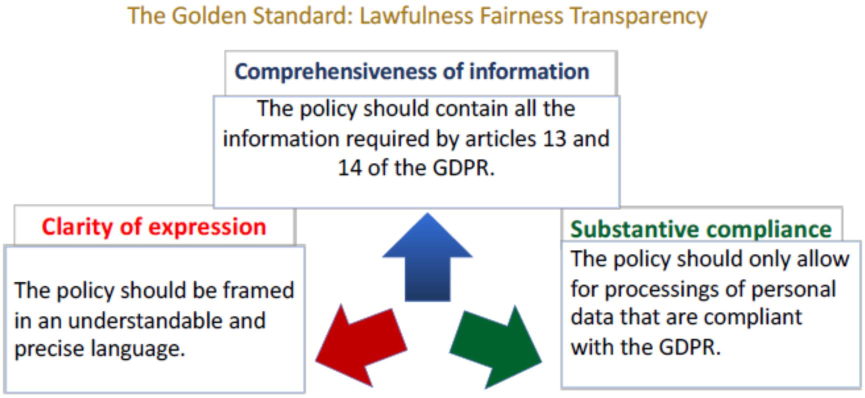
Human Legal experts are able to recognize unfair clauses thanks to their background knowledge. Rely on intuitions trained on past experience, provide reasons and use rationales to guide those.

Memory-Augmented NNs: process input and store information in some memory, understand pieces of knowledge relevant to a given query, retrieve concepts from memory and combine memory and query to make predictions.


Clause
b is not responsible or liable for the availability or accuracy of such Third-Party Services, or the content, products, or services available from such Third-Party Services"
Similarity
Aggregate
Yes
Keep reading?
No
Answer Module
Other
Unfair
KB
"The provider is not liable for any action taken from third parties or other people, including service and products, material and link posted by others"
...

After defining the KB and a list of legal rationales (possible explanations for unfairness), for each rationale define an identifier and link identifier to each clause marked as unfair in the dataset.

To classify a clause, the network makes a query to the knowledge base and compare with the list in the KB by similarity. Then the most relevant information (higher similarity score) is extracted and aggregated with the input clause (enhanced input) which will be used in another iteration of the same, so we look up into the memory multiple times.



The Golden Standard: Lawfulness Fairness Transparency

Comprehensiveness of information

The policy should contain all the information required by articles 13 and 14 of the GDPR.

Clarity of expression

The policy should be framed in an understandable and precise language.

Substantive compliance

The policy should only allow for processings of personal data that are compliant with the GDPR.

For each of this dimensions we defined a list of category we want to look for and different levels of achievement (optimal and suboptimal).

- 23 categories for **comprehensiveness of information** (identity of controller, contact details). Clauses where the categories of personal data are comprehensively specified and not vague are *fully informative* and *insufficient* in the other cases.

- 10 categories for **substantive compliance** (consent by using, advertising, licensing data, policy change, third party data transfers). **Policy**: When notice is given and new consent is required we have a *fair processing clause* while when notice is given but a new consent is not required we have a *problematic processing clause.* When no notice is given and a new consent is not required we have a clear *unfair processing clause.*

- 4 main indicators of vagueness for **clarity of expression**:

  1. **Conditional Terms**: the performance of a stated action or activity is dependent on a variable trigger *(depending, as necessary, as appropriate, as needed, sometimes)*
  2. **Generalization**: terms that vaguely abstract information practices using contexts that are unclear *(generally, mostly, commonly, usually, tipically, often)*
  3. **Modality**: includes modal verbs, adverbs, and non specific adjectives which create uncertainty w.r.t. actual action; whether an action is possible; does not include whether an action is permitted *(may, might, could, would, possible, possibly)*
  4. **Non-specific numeric quantifiers**: which creates ambiguity as to the actual measure *(certain, numerous, some, most, many, various, including, variety)*

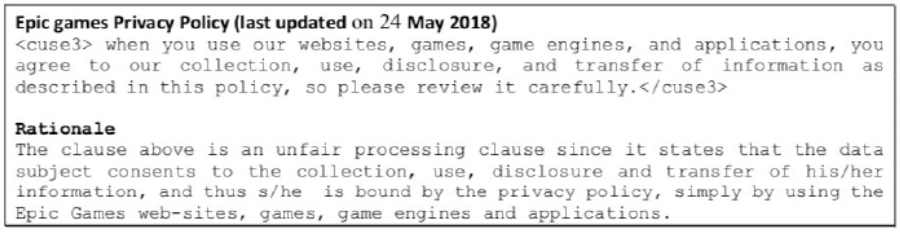
Web-Crawler: developed as a tool for automatic privacy policy monitoring. Two types of monitoring:

- Checking the date on the document

- Comparison of the content with the previously saved version.

At the moment, experimenting new method for privacy policies, multilingualism (a german version) and empower through transparency (linguistic transparency and provide explanations opening black box AI systems).

Failure under the substantive dimension



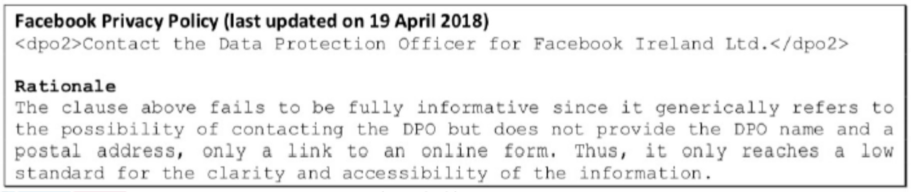
Epic games Privacy Policy (last updated on 24 May 2018)

<cuse3> when you use our websites, games, game engines, and applications, you agree to our collection, use, disclosure, and transfer of information as described in this policy, so please review it carefully.</cuse3>

Rationale

The clause above is an unfair processing clause since it states that the data subject consents to the collection, use, disclosure and transfer of his/her information, and thus s/he is bound by the privacy policy, simply by using the Epic Games web-sites, games, game engines and applications.

Failure under the comprehensiveness dimension



Facebook Privacy Policy (last updated on 19 April 2018)

<dpo2>Contact the Data Protection Officer for Facebook Ireland Ltd.</dpo2>

Rationale

The clause above fails to be fully informative since it generically refers to the possibility of contacting the DPO but does not provide the DPO name and a postal address, only a link to an online form. Thus, it only reaches a low standard for the clarity and accessibility of the information.