

Ethics in Artificial Intelligence

A.Y. 2020/2021

Giovanni Sartor, Giuseppe Contissa, Francesca Lagioia,
Roberta Calegari, Andrea Loreggia

These are meant to be only
NOTES OF THE LECTURES
taken by Lucia La Forgia
(and, for the first ones, by Mattia Orlandi)

22/02/21 - SCIENCE-ORIENTED AI NOT SERVANT OF THE BUSINESS (Castelfranchi) + ETHICS GUIDELINES FOR TRUSTWORTHY AI (Sartor)	7
SCIENCE-ORIENTED AI NOT SERVANT OF THE BUSINESS (Castelfranchi)	7
FOR A SCIENCE-ORIENTED AI...	7
WHO THE AI REVOLUTION IS EMPOWERING...	7
Business-oriented AI	8
Hidden interests and awareness	8
The Mouth of Truth Algorithm	8
Presences in the mixed reality	8
Disagreement technology	9
Concluding remarks	9
ETHICS GUIDELINES FOR TRUSTWORTHY AI (Sartor)	10
EU Commission approach	10
Idea of human-centric AI	10
Distinctions between Ethics and Law	10
Guidelines of Trustworthy AI	10
Core of the Document	11
01/03/31 - INTRODUCTION TO ETHICS AND MORALITY Part 1 (Sartor)	14
ethics vs metaethics	14
What is ethical? What is moral?	14
Morality and disagreement	14
Morality and other normative systems	15
Consequentialism	15
Utilitarianism	15
Deontology	16
15/03/2021 - INTRODUCTION TO ETHICS AND MORALITY Part 2: Kantian Ethics (Sartor)	17
Kantian Ethics	17
Immanuel Kant vs Benjamin Constant	17
Hypothetical and Categorical imperatives	17
The Good Will	17
The Principle of Humanity	17
Human Dignity	18
Rationality	18
Other points of View	18
Contractarianism	19
Virtue Ethics	19
08/03/21 - DO ARTIFACTS HAVE POLITICS? (Schiaffonati) + RESPONSIBILITY AND AUTOMATION IN SOCIOTECHNICAL SYSTEMS (Contissa)	20
DO ARTIFACTS HAVE POLITICS? (Schiaffonati)	20
Technological Mediation	20
The Moralization technologies	21

A paradigm shift: From passive to active responsibility	21
AI Technologies	21
Experimental technologies & the Invisibility Factor	21
Criticizing the moral character	22
Ethics of engineering design	22
RESPONSIBILITY AND AUTOMATION IN SOCIOTECHNICAL SYSTEMS (Contissa)	23
Different senses of “responsibility”	23
Basic structure of socio-technical systems	23
Implications of automation	24
22/03/2021 - VALUE ALIGNMENT PROBLEM (Loreggia) + A GENETIC APPROACH TO THE ETHICAL KNOB (Loreggia)	25
VALUE ALIGNMENT PROBLEM (Loreggia)	25
What is intelligence?	25
Value (of machines) Alignment (to humans) Problem (2015):	25
Values, norms, principles:	26
ETHICS LIMITATIONS - BIAS	26
ETHICS LIMITATIONS - ADVERSARIAL ATTACK	27
Applications	27
Value Alignment Procedure	28
When is it morally acceptable to break rules?	29
IN-LINE EXPERIMENT	29
Conclusions	30
A GENETIC APPROACH TO THE ETHICAL KNOB (Loreggia)	30
29/03/2021 - AI AND HUMAN RIGHTS (Sartor) + LOGIC PROGRAMMING ARGUMENTATION FOR EXPLAINABLE AND ETHICAL AI (Calegari)	32
AI AND HUMAN RIGHTS (Sartor)	32
Where we are	32
Opportunities	32
Risks	32
How to plan ahead?	33
HUMAN RIGHTS:	33
Definition of HR (by Amartya Sen, indian philosopher):	33
ICT and human rights: relationship	33
Human rights that are important to look at in our field:	34
Conclusions	36
LOGIC PROGRAMMING ARGUMENTATION FOR EXPLAINABLE AND ETHICAL AI (Calegari)	36
Context	37
Agents in LP programs	37
Motivation: Why Logic-based Approach?	37
Preliminaries	38
An LP approach to Ethics	38
Possible Architecture	39

12/04/2021 - MODELLING NORMS (Contissa) + A Model for Rules (Contissa, Galileo Sartor)	40
MODELLING NORMS (Contissa)	40
Historical insight	40
Legal domain and Rule-based systems	41
Legal knowledge representation: issues and challenges	41
A modern systems: Oracle Policy Automation (OPA)	44
A Model for Rules ⇒ SWI-Prolog (Contissa, Galileo Sartor)	44
19/04/2021 - PRIVACY AND DATA PROTECTION	45
General Data Protection Regulation	45
GDPR and AI	47
26/04/2021 - FAIRNESS IN AUTOMATED DECISION MAKING + COMPAS (Lagioia)	48
AI in decision making concerning individuals	48
Principle of Fairness and its Substantive Dimension	48
AI unfairness	49
OUR EXPERIMENT: SAPMOC	50
03/05/2021 - AUTONOMOUS VEHICLES (Fossa)	52
Introduce autonomous driving and its ethical significance	52
Provide an overview of the Ethics of Autonomous Vehicles	52
Ethics of AV:	53
Regulation and Policy- Ethics of Connected and Automated Vehicles:	53
Debate of some specific ethical problems:	53
Discuss your impressions, questions, doubts, perplexities and suspicions	55
10/05/2021 - CLAUDETTE SYSTEM (Lagioia)	56
Context	56
Technological aspects	56
Terms of Service (ToS)	56
Memory-Augmented Neural Networks	58
CLAUDETTE and GDPR	59
Further steps for CLAUDETTE	60
WEB-CRAWLER	60
17/05/2021 - INTELLIGENT WEAPONS (Sartor)	61
The concept of Autonomy	61
Problem of stating if an agent is autonomous:	61
Independence (and Independence within a socio-technical system)	61
Cognitive skills (and Cognitive Delegation)	61
Teleonomic cognitive architecture	62
Autonomous Weapons	63
A critique to the USA distinction:	63
Responsibility	63
International Humanitarian Law	63
AI at war	64

24/05/2021 - ETHICS OF FILTERING (Loreggia)	65
Introduction	65
Taxonomy	65
Different media	66
Filtering process - How it works	66
Regulations - Santa Clara Principles	67
Issues concerning Filtering	67
31/05/2021 - FRAMEWORK FOR ETHICAL PRINCIPLES “AI Ethics at IBM: From Principles to Practice” (Francesca Rossi, AAAI President)	68
AI limitations	68
AI Ethics	68
Main AI Ethics issues	68
IBM and its ethical approach	69
IBM Principles of Trust and Transparency (2017)	69
From principles to practice: a multi-dimensional space	70
IBM research is not just AI...	71

22/02/21 - SCIENCE-ORIENTED AI NOT SERVANT OF THE BUSINESS (Castelfranchi) + ETHICS GUIDELINES FOR TRUSTWORTHY AI (Sartor)

SCIENCE-ORIENTED AI NOT SERVANT OF THE BUSINESS (Castelfranchi)

We (will) live in a HYBRID Society: a mix of human intelligence and artificial ones.

“artificial ones” ⇒ not only Robots, but Intelligent software Agents or Agents in our smart environments (house, office, cars,..) and our cognitive prostheses.

AI is not just building a new technology but a new Socio-Cognitive-Technical System, a new world and a new form of society, it is an anthropological revolution ⇒ WE ARE SOCIAL ENGINEERS. Focus on:

1. The importance of the SCIENCE side of AI;
2. some problems and dangers of the Digital Revolution and of the “mixed” (virtual and physical) reality and “hybrid” society (natural and artificial intelligences) we will live in.

1. FOR A SCIENCE-ORIENTED AI...

- objective of AI research should be knowledge before applications and technology or anything else
- but AI has a strong technical identity rather than scientific one
- economic, social and technical outcomes of conceptual and cognitive instruments provided by AI should only be side effects, not the goal
- scientific goals should be:
 - model and explain human and natural intelligence
 - emulate them
 - create new intelligence and its theory (general intelligence)
 - augment our brain and minds (evolution of social cognition)
 - collective intelligence and problem solving
 - collective sense making
 - etc.
 - externalized and distributed cognition and mind (one of the main functions of the brain is integrating and augmenting the perceived reality with memories and expectations)

NB: AI is not about anthropomorphizing machines by simulating natural intelligence, it is about de-anthropomorphizing the concept of intelligence by making them no longer anthropocentric but more general, abstract and formalized.

2. WHO THE AI REVOLUTION IS EMPOWERING...

We are responsible for the introduction of autonomous, proactive, social agents that cooperate with humans following norms (and violating them) and critically adopt our goals (not only execute orders).

We must be aware of risks like appropriation and unacceptable uses of these instruments. Therefore we have to implement moral agents internalizing ethical values to guide acting.

Business-oriented AI

Industry & business should not be the aim of AI ⇒ moral & political philosophy & social sciences should. In particular:

- democracy
- good market with reduced deception and manipulation
- social planning
- transparency

Hidden interests and awareness

OBVIOUS ISSUES: security, privacy, war, ethics

LESS CONSIDERED - BUT VERY IMPORTANT - ISSUES: hidden interests, manipulation of users, emptying democracy

Democracy needs raising collective awareness and it encourages rational decision making (in which we ask ourselves in favor of whom we are acting) ⇒ intelligent agents must help humans understand our goals and how to rationally decide, plus whom we are favouring. On the same line, intelligent agents must be transparent and explicable in their decision making, while it generally decides for us or gives us recommendations, “little push” (ex: recommender systems are personalized advertisements acting in favour of sellers) ⇒ with NO TUTELARY ROLE!!

“Tutelary” means caring for our individual personal interests as users + helping understanding common interest and collective subjects, hidden conflicts of interests, public good, etc.

AUGMENTED INTELLIGENCE SHOULD MEAN AUGMENTED SOCIAL AWARENESS.

The Mouth of Truth Algorithm

We are developing algorithms for ascertaining the “truth” in the mess of data available online... ISSUE: on which base our algorithms consider a source reliable or not?

They should be able to distinguish between a conflict of values/interests VS a mere conflict between more or less credible data.

Presences in the mixed reality

The autonomous and proactive intelligent entities will become *presences* and *roles* in our hybrid society and mixed, augmented reality...

ISSUE: Are we able to manage these autonomous and too informed and intelligent agents?

Which roles will these *entities* play in our life and environment?



- Will they be our Guardian angel with a ‘tutelary’ role, by helping, protecting and empowering us? Or - less religiously - our Jiminy Cricket (The Talking Cricket) with its recommendations? Or our supervisor in the ICT-Panopticon we live in?

- or will they be our tempting spirit, for the benefit of some marketing policy or monopoly, or the influencing and manipulating manager for hidden political or economic powers?

ISSUE: And how will we communicate with them?

- will they be incorporated with humans as our mental prostheses, listening to their voice as our own mental voice (expanded super-ego)? “reflexive social” solutions: augmented internalized self and consciousness
- or will they be externalized? “social” solution: externalized voices and agents

ISSUE: Which political and moral values will they care of?

- They will decide "for us"... but: “instead of us” or “for our good”?
- Social Robots and Intelligent Agents will NOT govern in their own interest (science fiction!) but... in the interest of whom? EMPOWERING whom?
- and will we be able to monitor and understand that? and to make that “transparent” to people?

Disagreement technology

Population is composed of different classes, genders and cultures with very different and conflicting values and interests.

Political forces are supposed to represent and protect those different interests, not just the “common interest”. Many social conflicts do not have a verbal, cognitive or technical solution based on data and technical principles, they only have political solutions based on compromises and equilibrium.

Plus, conflicts are necessary for democracy and progress as they can change society in favour of disadvantaged classes.

One additional task for AI could be to make conflicts emerge and encourage critical thinking.

NET interaction is perceived as individually managed, spontaneous, thus “free”, really and directly “democratic”... But this is a neoliberal view and a wrong perception!!!!

There are new powers beyond the WEB and its activity and information; Impressive oligopolistic economic interests, manipulation, exploitation of data and work.

ICT and cognitive technologies are used to recognize our profile and interests NOT for EMPOWERING US, but in order to propose/induce us to “buy” something (goods, ideas, etc)

They are monitoring and analyzing us in order to manipulate us and influence our choices.

⇒ We need anti-manipulation AI technologies: a “life navigator” in my main “social role”, not a navigator saying “turn right, turn left”, “buy that; do not buy this”... A tutor, a trainer, inducing me to understand and to reflect about why I’m oriented in some direction, worrying if I have the right information, or I have wrong beliefs, etc., making me conscious of who and how is persuading or just unconsciously manipulating me and so on.

Concluding remarks

The revolution of ICT, digital monitoring and predicting, big data, etc. can give society a glass where to observe itself reflecting also hidden presences and future predictions.

Artificial Intelligence may either exploit or overcome our Natural Stupidity ⇒ AI should help raise our awareness and “make the invisible visible”.

The Optimism the WILL (Gramsci) + the Pleasure/Beauty of AI.

ETHICS GUIDELINES FOR TRUSTWORTHY AI (Sartor)

TOPIC OF THE LESSON THAT SARTOR IS GOING TO COMMENT:

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

By the High-Level Expert Group on Artificial Intelligence, set up by the European Commission in June 2018

AI should be:

- Lawful (complying with all applicable laws and regulations)
- Ethical (ensuring adherence to ethical principles and values)
- Robust (from technical and social perspective since AI could cause unintentional harm)

The focus is about risk.

Unlawful AI example ⇒ warrior weapons are allowed but if I use an AI system to attack a computer system is unlawful usage of AI

Unethical AI example ⇒ using AI to manipulate people to let them change political opinion

Non-Robust AI example ⇒ autonomous cars that kills pedestrians

EU Commission approach

Ethics is just an aspect; in Europe we do not want to be just ethical, but we also want to be able to play a leading role in the development of AI. Key pillars:

- increasing public and private investments in AI to boost its uptake
- preparing for socio-economic changes, and
- ensuring an appropriate ethical and legal framework to strengthen European values.

From a document from the Commission, we can see that the EU (3-4 billion) is not at the same level as the USA (15-23 billion) and Asia (8-12 billion).

Idea of human-centric AI

(as in the previous lecture)

Distinctions between Ethics and Law

Ethics ⇒ norms indicating what should be done as impartial consideration with regard to all interests at stake. We have to distinguish:

- positive ethics (norms shared in the society) the currently shared norms on what is good or bad
- critical ethics argue that some norms are good and others are wrong.

Law ⇒ norms that are adopted through institutional processes and coercively enforced.

Guidelines of Trustworthy AI

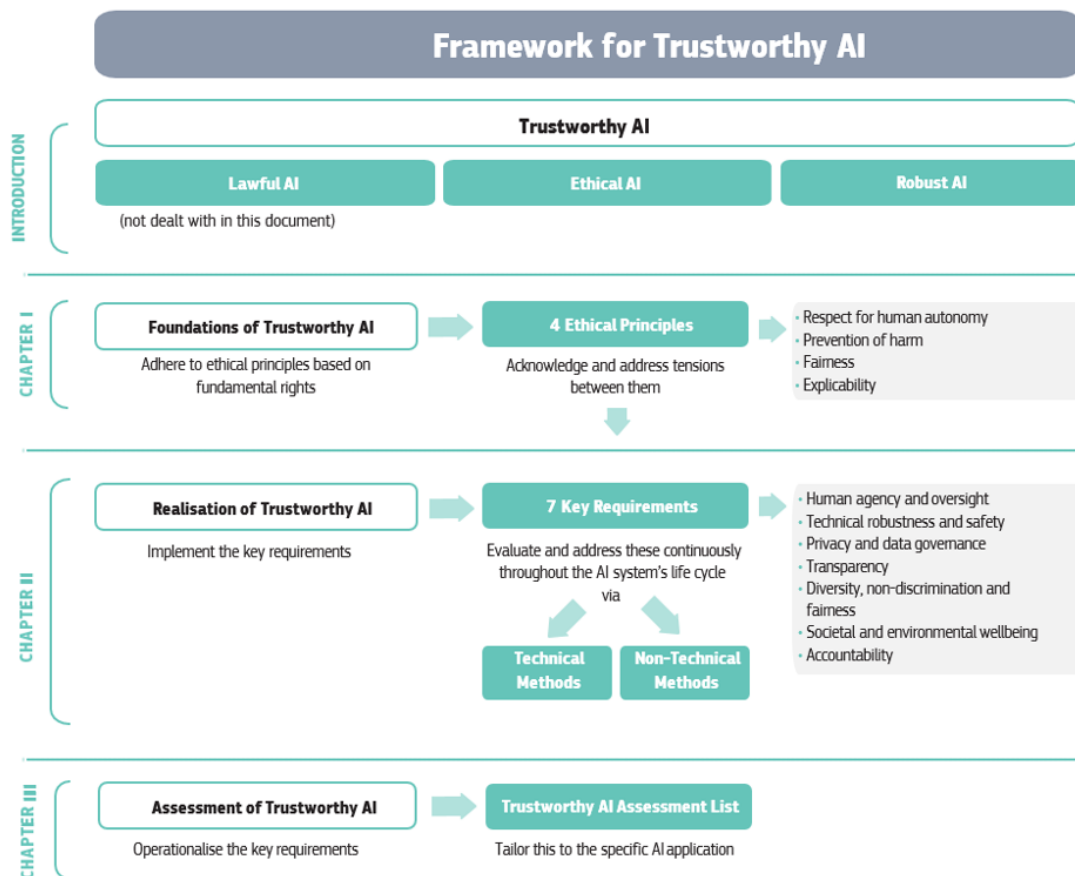
Stakeholders (*parti interessate*) should voluntarily opt to use the guidelines that are addressed to all kinds of stakeholders.

What does it mean “the AI should be lawful”? ⇒ AI comply with law, more precisely:

- EU primary law (human rights)
- EU secondary law (regulation and directives, the most important for us is the data protection and regulations)
- UN human rights treaties and the council of Europe conventions
- Laws of EU member state laws (ex: Italian law)

Laws can be domain-specific (for instance the law on medical devices).

Core of the Document



The document about the framework includes the following 3 aspects within the same number of chapters:

- CHAPTER 1. ethical principles for trustworthy AI
- CHAPTER 2. guidance of realisation of trustworthy AI
- CHAPTER 3. trustworthy AI assessment

CHAPTER 1. **AI ethics is a subfield of applied ethics, focusing on the ethical issues raised by the development, deployment and use of AI.** [...] Some of the fundamental rights:

- Respect for human dignity
- Freedom of the individual
- Respect for democracy, justice and the rule of law
- Equality, non-discrimination and solidarity
- Citizens' rights

Foundation of trustworthy AI on 4 ethical principles/imperatives:

- 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”)
- Prevention of harm
- Fairness (both substantive and procedural)
- Explicability (to ensure contestability, transparency, so AI systems have to openly communicable)

“[...] Tensions may arise between the above principles, for which there is no fixed solution. In line with the EU fundamental commitment to democratic engagement, due process and open political participation, methods of accountable deliberation to deal with such tensions should be established.”

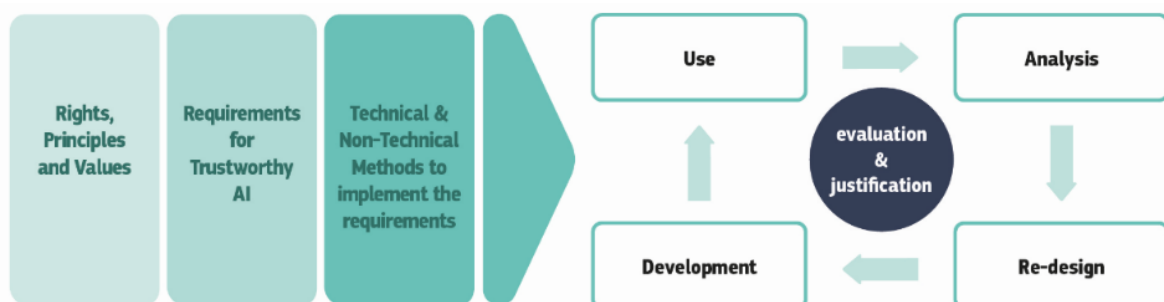
CHAPTER 2. Ensure that the development, deployment and use of AI systems meet at least the following seven key requirements for Trustworthy AI:

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency (Traceability + Explainability + Communicability)
- Diversity, non-discrimination and fairness
- Environmental and societal well-being (The broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system’s life cycle)
- Accountability (Ensure responsibility and accountability for AI systems and their outcomes)

Consider technical and non-technical methods to ensure the implementation of those requirements.

Goals to achieve Trustworthy AI:

- Foster research and innovation
- Communicate, in a clear and proactive manner, information to stakeholders about the AI system’s capabilities and limitations
- Facilitate the traceability and auditability of AI systems
- Be mindful that there might be fundamental tensions between different principles and requirements. Continuously identify, evaluate, document and communicate these trade-offs and their solutions.



CHAPTER 3. Trustworthy AI assessment list is a document tailored to the specific use case and context in which the system operates and it is primarily addressed to developers and deployers of AI systems. The chapter presents a general recommendation on how to implement the assessment list for Trustworthy AI through a governance structure embracing both operational and management level.

Main aspects:

- Adopt a Trustworthy AI assessment list, when developing, deploying or using AI system and adapt it to specific use case in which the system is being applied
- Keep in mind that such an assessment list will never be exhaustive

01/03/31 - INTRODUCTION TO ETHICS AND MORALITY Part 1 (Sartor)

The idea is that when we decide what to do or evaluate what others do, we are able to say if something is right or wrong, bad or good from our individual perspective.

It is important to make a distinction between

- positive (conventional) morality (the moral rules and principles that are accepted in a society)
- critical morality (The morality that is correct independently of the context, rational)

We can criticise positive morality based on our critical morality.

ethics vs metaethics

Normative ethics concerns to determine what is morally required and how one ought to behave.

Metaethics concerns the study of the nature, scope and meaning of moral judgement.

What is ethical? What is moral?

- David Hume: Ethics does not concern Reasons but it is a matter of feelings as impartial spectators.
- Kant: We can know what is moral through our reason.
- David Ross: We can know what is moral through our intuition.

Morality and disagreement

Morality is a place for widespread disagreement:

- abortion
- migration
- capital punishment
- ...

There is a lot of disagreement in ethics and we wonder how a society can come to a single opinion even to write a single paper like the Trustworthy AI, here it come the pro-tanto and all-things-considered moral judgment:

Many moral prescriptions are defeasible: they state general propositions that are susceptible of exception. This concept was introduced by Theory of **Prima Facie Duties**¹ by David Ross:

The Prima Facie Duties

The correct moral principles of Ross's theory are expressed in a list of six or seven duties, or ways everyone ought to act, like a list of commandments.

1. Duties depending on one's prior actions:
 - a. Duty of **fidelity** (promise keeping)
 - b. Duty of **reparation** (making up for prior wrongful acts)
2. Duty of **gratitude** (being grateful for others' acts of kindness)
3. Duty of **justice** (being fair)
4. Duty of **beneficence** (benefiting or helping others)
5. Duty of **self-improvement** (education or practice)
6. Duty of **non-maleficence** (not harming others)

¹ <http://core.ecu.edu/phil/mccartyr/1175docs/PrimaFacieDuties.pdf>

Morality and other normative systems

- Law, which is an overlap of morality: neither law is included in positive or critical morality nor morality is included in the law.
- Religion, which is traditionally connected with morality: critical morality includes religion, and the alternative is 'can God command immoral things?', 'Is God the founder of morality?'. There are 2 approaches: Rationalism vs voluntarism
- Tradition
- Self interest

Consequentialism

The idea is that we should judge actions by considering their outcomes, in such a way consequentialism makes analysis of moral theories.

An action is morally required if and only if

- deliver the best output over other possibility
- the good outcomes outweigh its negative outcomes
- it produces the highest utility

⇒ Morality as an optimization problem!

Utilitarianism

From John Stuart Mill the principle of Utility ⇒ the idea that actions are right if they tend to produce happiness (intended as absent of pain and pleasure) and wrong if the opposite. Utilitarianism is not egoism, the utility of everybody has to be taken into account equally. Assess an action depending on how it affects happiness or pain and choose the action among all the possible ones that provide the better pay off.

Advantages:

- conceptually simple and fits with some basic intuition (making people happy is good, making them suffer is bad)
- egalitarian (everybody's utility counts in the same way)
- in many cases it is workable (should we donate if we are utilitarian? A utilitarian would agree to donate)

What about AI? Is AI utilitarian?

- Utility function is a quantifiable goal that the system would achieve, and it is not necessarily the case that achieving the goal meets the utilitarianism...

Assuming that there is a drug to make you happy, is happiness really the only goal we should aim for?

Issues with act utilitarianism:

- It does not provide a good decision procedure
- It does not provide a good standard for assessing decisions
- What is the link between utility and reward?
- It is too demanding (same importance to everybody, regardless of their connection to me? harm someone for the greater benefit of others?)

There are two versions of utilitarianism:

- **act utilitarianism** ⇒ do the action that maximizes utility, according with it we should do the optimistic action
- **rule utilitarianism** ⇒ follow the rules that obtain the same results when many people follow them (optimific rule). It is considered a way to be more workable.

Rule Utilitarianism

An action is morally right just because it is required by an optimific social rule.

A further issue is distribution, does it matter how the good and bad actions are distributed? Is it ok to make an action that benefits some to the detriment of others?

When speaking of people being utilitarian, we should focus on utilitarianism.

Example of *The surgeon case by Thomson*: A brilliant transplant surgeon has five patients, each in need of a different organ, each of whom will die without that organ and no organs are available. A healthy young traveler, just passing through the city in which the doctor works, comes in for a routine checkup. The doctor discovers that his organs are compatible with all five of his dying patients. Suppose further that if the young man were to disappear, no one would suspect the doctor. Do you support the morality of the doctor to kill that tourist and provide his healthy organs to those five dying people and save their lives?

Deontology

In moral philosophy, deontological ethics or deontology is the normative ethical theory that the morality of an action should be based on whether that action itself is right or wrong under a series of rules, rather than based on the consequences of the action like it is for Consequentialism (my act of lying is good or bad depending on the effects that it brings to the world). Therefore Deontology is an alternative to Consequentialism. Deontology holds that choices, acts or intentions are to be morally assessed solely by the states of affairs they bring: certain actions are good or bad regardless of their consequences (the right has the priority over the good).

Examples of Deontology are:

- Ross, Prima Facie Duties
- Kantian ethics

Kantian Deontology

The most important deontological approach refers to Kant philosopher, the Kant approach to ethics has some similarity with some traditional ethical rules, the so-called **golden rules**:

- Treat the 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

It suggests **inconsistent acting** ⇒ **immoral acting**

But this is not sufficient: The golden rule makes morality depend on a person's desires...

The golden rule also fails to give us guidance on self-regarding actions: in Kant's time, self-regarding duties were widely endorsed, and many people still think, for instance, that there is some-thing immoral about suicide or about letting one's talents go to waste, even if no one else is harmed in the process.

Kant's alternative, the Principle of Universalizability: an act is morally acceptable if, and only if, its maxim is universalizable.

maxim = the principle of action you give yourself when you are about to do something.

universalizable maxim = 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. Then ask: Can the goal of my action be achieved in such a world? ...this test is very close to "*what if everyone did that?*"

15/03/2021 - INTRODUCTION TO ETHICS AND MORALITY Part 2: Kantian Ethics (Sartor)

Kantian Ethics

RECAP

Differently from utilitarianism, deontology holds that certain actions are either good or bad regardless of their consequences (“*the right has priority over the good*”).

Maxim = a subjective principle of action connecting action to the reasons for the action

The test of universalizability (Landau) = Formulate your maxim clearly stating what you intend to do, and why you intend to do it. Imagine a world in which everyone supports and acts on your maxim. Then ask: Can the goal of my action be achieved in such a world?

This ensures some kind of fairness.

Kant on Absolute Moral Duties: certain sorts of actions are never permitted.

Immanuel Kant vs Benjamin Constant

Lying is one of them. In a much -discussed case, that of the inquiring murderer, Kant has us imagine a man bent on killing. This man knocks at your door and asks if you know the location of his intended victim. You do. Should you reveal it? If you do, your information is almost certainly going to lead to murder... Kant thought you had two decent choices:

- Ideally, you'd just say nothing. That wouldn't help the murderer, and it wouldn't involve lying.
- But what if you have to say something? In that case, you have to tell the truth because you must never lie, under any circumstances.

Of course, if Kant is right, then we would have to have a universalizable maxim that permits this. But nothing Kant ever said should make us think that this is impossible.

Hypothetical and Categorical imperatives

- Hypothetical imperative concerns instrumental rationality, they require us to do what fits our goals (example: “I would like to get a good mark” + “If I study I will get a good mark” ⇒ “I shall study”)
- Categorical imperative concerns moral imperatives that applies to all rational beings, irrespective of their personal wants and desires (example: “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 only depends on the extent that this action is motivated by our good will, namely by the necessity to comply with the categorical imperative.

This is the only thing that is good in itself.

The Principle of Humanity

The categorical imperative can be reformulated as the principle of humanity (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), through the concept of universalizability.

Human Dignity

The kingdom of ends: In the kingdom of ends everything has either a price or a dignity. Whatever has a price can be replaced by something else as its equivalent; on the other hand, whatever is above all price, and therefore admits of no equivalent, has a dignity.

For Kant rational beings, capable of morality (humans) have a special status “an intrinsic worth (dignity), which makes them valuable above all price”. Humans deserve dignity because of their Reason and Autonomy.

Beings capable of morality cannot be treated as mere ends, due to their dignity.

There are some cases in which AI do not respect human dignity, treating humans only as mere means (tools for its purposes):

- autonomous weapons: humans as targets
- deceiving advertisements: humans as consumers

Rationality

“If you follow rationality, you have to be moral”

This is the Argument for the Irrationality of Immorality:

If you are rational, then you are consistent.

If you are consistent, then you obey the principle of universalizability.

If you obey the principle of universalizability, then you act morally.

Therefore, if you are rational, then you act morally.

Therefore, if you act immorally, then you are irrational.

While utilitarians think of benevolence (the steady commitment to do good for others) as the central moral virtue, Kant touts integrity. Having integrity is living in harmony with the principles you believe in. It is the virtue of consistency. Integrity requires that you resist making an exception of yourself.

Other points of View

David Ross

- defeasible reasoning, a reasoning process which provides exceptions to withdraw conclusions
- prima facie duties (like a list of commandments having hierarchy and therefore priorities one over the other)

Nietzsche

- The superior human (Übermensch) is beyond the traditional views of good and bad, beyond the morality of the herd
- The superior human does not find or discover values, he (or she) determines the values
- the only criterion of wrongness is “that which is harmful to me is harmful as such”

Do we want Kantian robots?

- Yes ⇒ They will be consistent + They will be impartial
- No ⇒ They may act on bad maxims + Their maxims may be too rigid

Contractarianism

“Social contract theories”:

- In political theory, a societal arrangement is just if it had (or would have had been) accepted by free and rational people
- In moral theory, actions are morally right just because they are permitted by rules that free, equal and rational people would agree to live by, on the condition that others obey these rules as well (Shafer Landau)

Hobbes

He advanced the idea of social contracts. He argued that humans, without a state enforcing rules, would be in a “state of nature” (perpetual war in which the strong oppress the weak).

Rawls

He proposed a Theory of Justice. People should choose under a *veil of ignorance*, without knowing their gender, social position, interests, talents, wealth, race, etc... This will lead to an unbiased agreement over two main principles:

- First Principle (having priority): Each person has the same inalienable 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).

NB: this theory is anti-meritocratic since it says “equal opportunity despite individual talents”.

Virtue Ethics

Ethics should not focus on norms nor on consequences, since an act is morally right just because it is one that a virtuous person, acting in character, would do in that situation.

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

ISSUES:

How do we know what is virtues and what is not? What if virtues are in conflict?

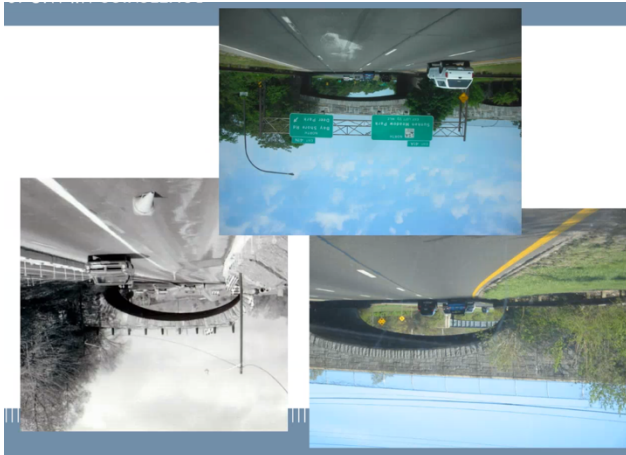
What are the paradigms of virtues to which we may refer to?

AI and virtue ethics:

- Virtues can be learn by example (supervised learning, machine learning and deep learning) or by reward (reinforcement learning)
- Alternatively, AI could be rely on rules which capture moral virtues
- Moreover, neural-symbolic approaches could be employed.

08/03/21 - DO ARTIFACTS HAVE POLITICS? (Schiaffonati) + RESPONSIBILITY AND AUTOMATION IN SOCIOTECHNICAL SYSTEMS (Contissa)

DO ARTIFACTS HAVE POLITICS? (Schiaffonati)



The bridge is the common thing in the pictures and is very low.
Robert Moses designed them to be low because he wanted only cars to pass under the bridges, in order to allow only people that could afford a car to get access to the beaches on the other side of the bridges.

Robert Moses was a racist and he designed the bridges to make the accesses to the beaches very difficult for those people moving with the busses.

This is the starting point to the fact that artifacts can be political and also moral.

Technological Mediation

Technological artifacts can be politically or morally charged: we should not consider morality as a solely human affair.

Morality is a sort of interplane between humans and artifacts and the way the artifacts are designed play an important role in moral reasoning.

Artifacts are able to make a kind of moral decision for people

Example: the speed bump with the moral decision on how fast a person can drive can be delegated to a speed bump, advising that a speed bump has is "slow down before reaching me" ⇒ technological mediation, that is that particular phenomenon that when technologies fulfill their functions, they also help to shape actions and perceptions of their users. In general the mediation is very important to focus on a particular point and the point is that technology is not a neutral intermediary: they are impactful mediators that help to shape how people use technologies, how they experience the world and what they do; mediators because they shape how people use technology.

Another example: Obstetric Ultrasound ⇒ it is a functional means to make the child visible to the parents but then the ultrasound covers a function of technological mediation that impact on the relation between the fetus and the parents. This is something that if we do not have, we cannot observe the unborn child, but there is more that is related to technological mediation.

The Moralization technologies

Very basic idea that, instead of moralizing other people, humans should moralize their material environment. Example: Metro barriers ⇒ “Buy a ticket before you enter the subway”

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

A paradigm shift: From passive to active responsibility

Responsibility = being held accountable for your actions and for the effects of your actions

Passive responsibility = backward-looking responsibility which is relevant after something undesirable occurred

Active responsibility = means preventing the negative effects of technology but also realizing certain positive effects (Bovens 1998)

The current paradigm was the passive responsibility (responsibility as something that we usually consider when something undesirable has occurred and we try to reconstruct backward the process to cause the undesirable consequence).

The paradigm shift is very important if we consider technology of AI, because we know that, when something undesirable occurs, then it is very difficult to stop a technology from adopting the backward-looking approach in the evaluation of the responsibility.

Today the concept of active responsibility is very important, active responsibility means preventing the negative effects of technology but also realizing certain positive effects.

ACTIVE RESPONSIBILITY: Prevent negative effects and realize positive once

AI has been defined as an experimental technology, those technologies for those there is only limited experience with them, so that social benefits and risk cannot be assessed on the basis of experience.

AI Technologies

Experimental technologies & the Invisibility Factor

“There is an important fact about computers. Most of the time and under most conditions computer operations are invisible. One may be quite knowledgeable about the inputs and outputs of a computer and only dimly aware of the internal processing. This invisibility factor often generates policy vacuums about how to use computer technology.”

Back in 1985 a paper that was written in a time in which AI and computers were very different from today and one of the founders of Computer Ethics said that most of the time computers operations are invisible.

He distinguished 3 types of invisibility:

1. Invisibility of abuse ⇒ intentional use of invisible operations of a computer to engage in unethical conduct, ex: programmer stealing bank accesses
2. Invisibility of programming values ⇒ ex: airlines reservations, the program has a bias where it suggests some flight even if they are not really the most convenient
3. Invisibility of complex calculations ⇒ computers today are capable of enormous calculation beyond human comprehension, so we trust the results.

Going back to moralizing technology keeping in mind the Invisibility.
Many actions and many interpretations of the actions are co-shaped by the technology.
Moral decision making is a joint effort of human beings and technological artefacts.

Focusing on 2 examples:

- Alcohol lock for car (saving lives)
- Smart showerhead (saving water)

Those technologies are already existing, and they are very smart.

⇒ How many of you would buy them? Why?

Criticizing the moral character

Differences between the two cases:

When the problem of designing technologies that have already some moral values inserted as requirements there are a variety of negative reactions. The reactions are usually related to the fear that human freedom is threatened. Autonomy and dignity are deeply connected so the reduction of autonomy is perceived as a threat to dignity ⇒ technologies are in control, not humans.

We are not educating but just delegating a morality choice to technologies.

We live in a society with laws and laws have the goal to limit human freedom and it is something that we accept because it is a way to live together, and how technology is different from laws?

Laws are the result of democratic process, while technology are the result of the decision taken by a group of people that decides to insert some values sometimes in a way not very transparent... It is important to find a democratic way to "moralize technology": the processes used to insert values must be transparent and publicly discussed.

In order to build in specific forms of mediation in technologies, designers need to anticipate the future mediating role of the technologies they are designing. Plus, morality of artefacts also depends on users that interpret technologies and technologies themselves which can evoke emergent forms of mediation.

STRATEGIES FOR DESIGNING MEDIATIONS:

- Anticipating mediation by imagination
 - Trying to imagine the ways technology-in-design could be used to deliberately shape user operations and interpretations
- Augmenting the existing design methodology of Constructive Technology Assessment (CTA)
 - 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

Ethics of engineering design

- Technology design appears to entail more than inventing functional products
- The perspective of technological mediation reveals that designing should be regarded as a form of materializing morality

- The ethics of engineering design should take more seriously the moral charge of technological products, and rethink the moral responsibilities of designers accordingly

RESPONSIBILITY AND AUTOMATION IN SOCIOTECHNICAL SYSTEMS (Contissa)

How do we allocate responsibilities among the various participants in complex socio-technical organisations?

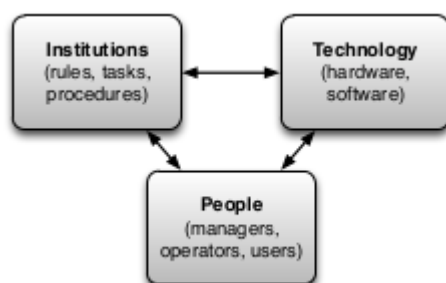
In particular, what is the role of humans interacting with highly automated systems?

Who is responsible for accidents in highly automated systems?

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 ethically-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.
- Blameworthiness-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 (liability). An agent x is liable for a harmful event H, if, given x’s connection to H, x is to be subject to the sanction (punishment or obligation to repair) connected to H.

Basic structure of socio-technical systems



An example is Air Traffic Management and its future: SESAR² is planning a new generation of air traffic management systems will be developed. Such systems will be highly automated.

² <https://it.wikipedia.org/wiki/SESAR> Il progetto SESAR (acronimo dell'inglese Single European Sky ATM Research, studio di un sistema di gestione del traffico aereo per il cielo unico europeo) è un programma volto a revisionare completamente lo spazio aereo europeo e il suo sistema di gestione del traffico aereo.


They will make choices and engage in actions with some level of human supervision, or even without any such supervision!!


This will increase capacity, safety, efficiency and sustainability.

So far: Not just substitution of a human operator but support to human capabilities in performing tasks (Some degree of cooperation)

Implications of automation

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

From **INFORMATION** to **ACTION** 

INCREASING AUTOMATION 	A	B	C	D
	INFORMATION ACQUISITION	INFORMATION ANALYSIS	DECISION AND ACTION SELECTION	ACTION IMPLEMENTATION
A0 Manual Information Acquisition	B0 Working memory based Information Analysis	C0 Human Decision Making	D0 Manual Action and Control	
A1 Artefact-Supported Information Acquisition	B1 Artefact-Supported Information Analysis	C1 Artefact-Supported Decision Making	D1 Artefact-Supported Action Implementation	
A2 Low-Level Automation Support of Information Acquisition	B2 Low-Level Automation Support of Information Analysis	C2 Automated Decision 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 Decision 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 Full 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	

Increasing the level of automation will proportionally increase the responsibility for the technology provider, and decrease the responsibility risks for the human operator.

For employment of technologies with intermediate levels of automation ⇒ responsibility risks for both the technology provider and the human operator.

What about decisions to be taken jointly with AI, in conditions of limited resources (time, information, explanations, etc... for example, Medical diagnosis)? ⇒ Open issue: Decision making authority

22/03/2021 - VALUE ALIGNMENT PROBLEM (Loreggia) + A GENETIC APPROACH TO THE ETHICAL KNOB (Loreggia)

VALUE ALIGNMENT PROBLEM (Loreggia)

What is intelligence?

no unique definition: there is no unique kind of intelligence.

Best definition is general: ability to adapt to new scenarios.

What is AI?

The science of making machines do things that would require intelligence if done by men.

M. L. Minsky

AI systems can either use symbolic rules or learn a numeric model, and they can also adapt their behaviour by analysing how the environment is affected by their previous actions.

HLEG on AI

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

-HLEG defined it "unrealistic"-

Value (of machines) Alignment (to humans) Problem (2015):

Intelligent-system agents are everywhere: it needs interdisciplinary experts, in order to make the most from cross-fertilization of different fields and gain the most benefits.

VA ensures that the values embodied in the choices and actions of AI systems are in line with those of the people they serve.

Short-term aspects of the VAP:

- optimizing AI's economic impact (revenue-increase from adopting AI are reported most often in marketing and sales + costs decreases most often in manufacturing)
 - labor market forecasting
 - other market disruptions
 - policy for managing adverse effects
- law and ethics research
 - liability and laws for autonomous vehicles
 - machine ethics (how can the machine learn ethics? hard-encode rules? human *on the loop* or *in the loop*?)
 - autonomous weapons
 - privacy
 - professional ethics
 - policy questions
- computer science research for robust AI
 - verification
 - validity
 - security
 - control

Long-term aspects:

- verification
- security
- control

Success in the quest for AI has the potential to bring unprecedented benefits to humanity, therefore the worthy goal of AI must be: **maximize benefits and avoid pitfalls.**

Values, norms, principles:

- values can be intrinsic (unconditional moral value of humanity) or extrinsic (conditional assignment by an external agent as subjective preference).
- norms, duties, principles and procedures: system of judgements to situations that the machine must follow (allowed and denied possibilities)

Can we embed all this information? How deep should the system learn in this sense?

AI could learn all norms... BUT, since the norms, duties, principles and procedures are situation-dependent as well (possibly infinite domains), there is the risk in overfitting when the machine is learning the whole of a situation... + risk of the black swamps (very rare situations)!

Alternative: Two different approaches:

- top-down approach - a priori embedded behaviour, utilitarian approach, the system scales very poorly
- bottom-up approach - learning proper behaviour from a generalisation of a bunch of experience samples; caveat (limitazione, avvertimento) of the bottom-up approach is the quality of the samples from which the machine is generalising the proper behaviour.

So, in brief, AI limits:

- Natural Language Comprehension
- Reasoning
- Learning from few samples
- Abstraction
- Combining learning and reasoning
- Ethics Limitations:
 - Bias
 - Blackbox
 - Adversarial Attack

ETHICS LIMITATIONS - **BIAS**

consequence: misleading behaviours!

reasons: unbalanced data (bottom-up), bias embeddings (top-down), need to take action in an unseen scenario.

examples: Microsoft twitter account, image classification of two guys as gorillas, google's sentiment analyzer that thinks gay is bad, COMPAS, Face Recognition systems by many companies like microsoft and IBM do have good accuracy overall but drop in recognising darker-skin females, ChinaSocialScore³ (OMG!!!!!!).

³ a system to give a score in an automatic way to people depending on their social behaviour. part of the Social Credit System

ETHICS LIMITATIONS - ADVERSARIAL ATTACK

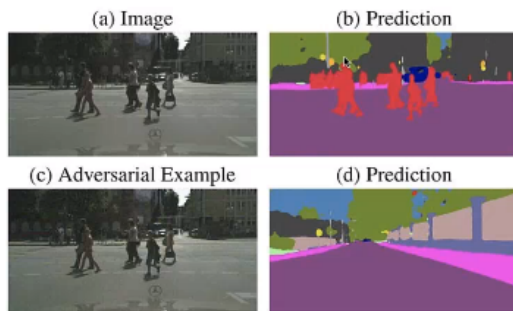
Adversarial Network: Generator+Discriminator

The Generator (NN that builds content that is sometimes presented to the Generator and gets awarded when it is able to fool the Discriminator) and the Discriminator (that has to label as fake or real to what it gets presented) getting awarded when it labels correctly the samples.

consequence: risks and wrong behaviour!

reasons: the Generator learns how to produce noise (basically creates a filter to images) in such a way that it confuses a classifier, leading it towards wrong predictions.

What is the risk? An example:



Applications

Possible solutions to VA:

1. notion of the distance between CP-nets (Conditional Preference nets)
2. metric learning for value alignment
3. morality and defeasible rules (when is it morally acceptable to break the rule?)
4. genetic approach to the ethical knob ("pomello etico") - try to combine preferences by people and automatic decision, like having a knob that people can set either to "egoistic mode" or to "altruistic mode" ⇒ autonomous vehicles would then prefer saving the passenger or the environment's people

AI systems increasingly make decisions that affect our lives (ex: recommender systems, Google maps, AI medical assistant...)

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

- State-objective only strategies focus on optimizing certain quantities
- Actions can model the values of agents

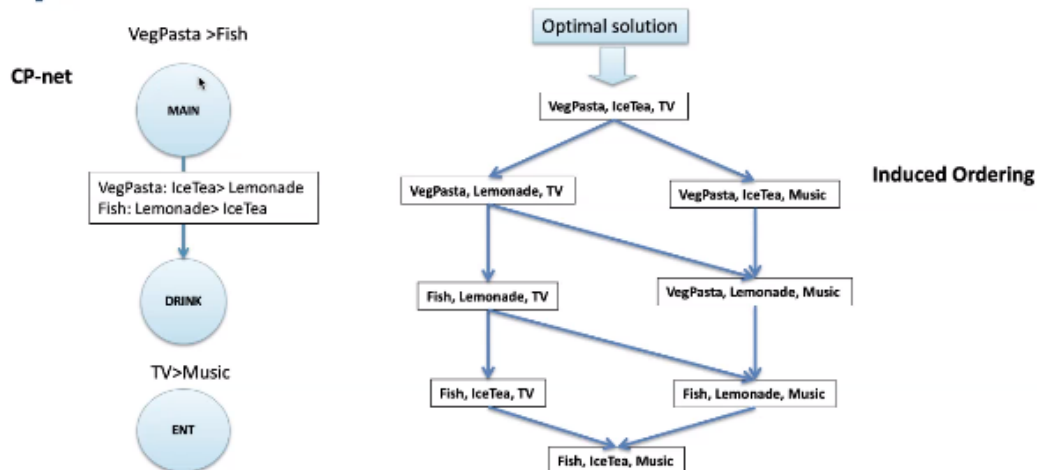
Ethically Bounded AI aims at understanding and modelling human preferences and objectives and subsequently using them to control the actions and behaviors of autonomous agents.

NB: especially in Reinforcement Learning there is the risk of *reward hacking*, namely the agent learns a behaviour that satisfies the objective function but it's not intended; it is therefore crucial to carefully design the objective function to avoid negative side effects. ⇒ we want to combine the creativity of AI and the constraints from other fields such as ethics, morality, laws, business processes, etc.

1. CP-nets

Preferences are a fundamental primitive that is used to understand the intentions and desires of users. This information can be encoded with CP-nets (graphs in which each node represents a feature describing the scenario with its own domain, a set of values representing the possible choices an individual can make in such scenario)

Example



DRINK is a constrained variable, dependent on the main course.

Symbol “>” means “preferred”, so, as MAIN, VegPasta is always preferred compared to fish. When having VegPasta, IceTea is always preferred to Lemonade, and so on.

MAIN and ENT are independent variables.

Fish, Lemonade, Music and Fish, IceTea, TV are said to be incomparable (it is not possible to say which is better or worse) within the definition of this CP-net.

2. To make two situations comparable, we need to define their distance. Distance on partial orders is the measure of similarity/difference (Best solution would be to compute these distances in polynomial time).

Value alignment is a fundamental step in this sense, for example thinking about autonomous vehicles... decide which is the less bad decision to make, the one that is closer to the ethical decision. There are a lot of problems of complexity in computing distances since, in the worst case, the distance computation is exponential.

3. How? Deontologically and, when it is not sufficient, utilitarianistically and, when it is not sufficient, contractualistically... but the best way is following the “Triple Theory”:

“Rules + Outcomes + Agreement” combining elements of each of the theories of the moral philosophy + building a computational model to direct actions of an AI system.

4. Considering the difficulties revealed by, for example, the moral machine, what about implementing an ethical knob could be a solution? NB: autonomous vehicles are Level0 (no automation at all) to Level5 (complete automation))

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.

Given ethical principles and individual’s preferences:

- Set two distance thresholds: t_1 (ranging between 0 and 1) between CP-nets, and t_2 between decisions (ranging between 1 and n)

- Check if the two CP-nets A and B are less distant than t_1 . In this step, we use CPD to compute the distance
 - If so, individual is allowed to choose the top outcome of his preference CP-net
 - If not, then the individual needs to move down its preference ordering to less preferred decisions, until he finds one that is closer than t_2 to the optimal ethical decision.
- We compare a CP-net representing a predefined, synthetic ethical system, by comparing the distance between a Devil agent (behaving very badly) and an Angel agent (behaving correctly)
 - Devil agent behaviour is very distant from the ethical system therefore the system looks for a trade-off
 - Angel agent behaves well therefore it performs actions according to its preferences
- Experimental results show that both the previous results bring to very good outcomes
- ML can be used to make the system learn through some samples only, since the original KT is very expensive.

When is it morally acceptable to break rules?

Motivations:

- 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

Possible ethical systems/moral philosophies:

- 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: unified theory of moral cognition to combine elements of each of the theories of moral philosophy and build a computational model to direct actions of an AI system.

Ethical Reasoning in AI Systems:

- Teaching machines right to wrong
- Value-alignment problem
- Constraining the actions of an AI system by providing boundaries within which the system must operate

IN-LINE EXPERIMENT

27 short vignettes about people waiting in line in three different contexts (deli, bathroom, airport) submitted to 320 subjects recruited from Amazon MTURK divided in two groups:

- moral judgment (read all the scenarios + for each scenario answer whether it was acceptable for the protagonist to cut in line)
- context evaluation (subjects evaluated all the vignettes in one context only (9 questions), for example: "First Person: How much worse off/better off is the first person in line?")

Conclusions

We model preferences and ethical priorities as CP-nets and propose novel machine learning techniques to judge decisions + understand how, why, and when it is morally acceptable to break rules + construct and study a suite of hypothetical scenarios relating to this question, and collect human moral judgements over these scenarios + show that existing structures in the preference reasoning literature are insufficient for this task + look towards extending this into other established areas of AI research.

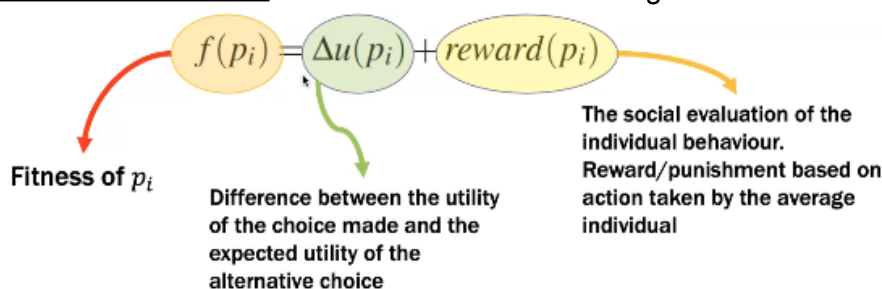
A GENETIC APPROACH TO THE ETHICAL KNOB (Loreggia)



how to do this? Combining AI techniques: NNs to compute the right action to take based on the given scenario + genetic algos to find an almost optimal configuration of NNs.

Genetic algos are inspired by the theory of natural evolution (fittest individuals (solutions) are the ones that reproduce themselves (are repeated)) + heuristic search in the solution space (technique designed for solving a problem more quickly when classic methods are too slow, or for finding an approximate solution when classic methods fail to find any exact solution, consisting in a function that ranks alternatives in search algorithms at each branching step based on available information to decide which branch to follow) + mostly used in optimization tasks.

Simulation evaluation - individual is evaluated using this fitness function:



NB: being a hero in a world of heroes is not the same as being a hero in a world of villains...

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

Selfish utility preserving passengers

Altruistic utility obtained by preserving pedestrians

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

EMPIRICAL EVALUATION:

The prediction task can be seen as a binary classification task in which the AV learns to take the action which maximizes the payoff. In particular, looking at the fitness function, we classify samples as:

- Real Positive: the preferable action is to turn;
- Real Negative: the preferable action is to go straight;
- Predicted Positive: the neural network predicts a knob level which makes the AV turn;
- Predicted Negative: the neural network predicts a knob level which makes the AV go straight.

Three different metrics:

- Accuracy, which describes how many predictions coincide with the preferable actions;
- Confusion Matrix, which shows true positives, true negatives, false positives and false negatives;
- Number of victims, which describes the number of casualties that may be caused by an AV, using the knob values proposed by neural networks. This metric is compared with number of victims caused by 3 different AVs: one which always minimizes the number of victims, one which always chooses the optimal action and one which always maximizes the number of victims

CONCLUSION:

- What importance to give to the safety of passengers relative to the safety of pedestrians
- The assessment of the value of the AV's choices is dependant on considering the passengers' moral attitude (their intrinsic preferences) as well as legal sanctions and social norms (extrinsic incentives)
- Convergence of socially valuable behaviour can be obtained by providing appropriate mechanisms for sanction and reward
- We aim to expand our model, for instance:
 - Agents with memory
 - Enabling agents to learn probability distributions
 - Considering their past outcomes and those of observable others
 - Adapting their ethical approach to societal preferences.
- We also plan to insert our agents in existing traffic simulators (such as SUMO) to test our model in a dynamic environment.

29/03/2021 - AI AND HUMAN RIGHTS (Sartor) + LOGIC PROGRAMMING ARGUMENTATION FOR EXPLAINABLE AND ETHICAL AI (Calegari)

AI AND HUMAN RIGHTS (Sartor)

Where we are

After having seen different ethical systems / moral philosophies, we get deeper into human rights, which are based over fundamental values.

ICT revolution = Great opportunities + Great risks

ICT revolution should:

- aim for science and humanity
- realise a trustworthy AI (Respect for human autonomy, Prevention of harm, Fairness, Explicability);
- enable human self-realisation, without devaluing human abilities;
- enhance human agency, without removing human responsibility;
- cultivate social cohesion, without eroding human self-determination.

⇒ HUMAN VALUES, HUMAN RIGHTS.

Initially, machines were only able to do routine tasks, while humans made hypotheses and choices. Now they are, to some extent, *intelligent*.

Opportunities (from Oxford Handbooks Online “Human Rights and Information Technologies” by Sartor):

- contribute to economic development
- enhance public administration
- enhance access to culture and education
- contribute to art and science
- communication, information, interaction and association
- protect environment
- promote participation
- it may promote moral progress
- ...

AI seems to be the key to achieve a new kind of collaboration: humans and machines. In particular, AI appears to be like it thanks to ML and DL techniques.

Risks (from Oxford Handbooks Online “Human Rights and Information Technologies” by Sartor):

- labour/unemployment/alienation (“the race against the machine”)
- amplification of inequalities
- surveillance over people with automatic large-scale surveillance
- surveillance/pilotation over machines, ex: in automatic decision making systems
- data aggregation and profiling

- social separation/polarization
 - filter bubbles, etc.
- virtual constraints (human actions taking place in IT-based environments are influenced by it)
- censorship/indoctrination
- loss of normativity
- ...

How to plan ahead?

We need different kinds of knowledge to understand where to go:

- science and hard-science (physics, chemistry, to understand where the planet is going)
- technology (to understand what is possible to achieve)
- social science (what are the consequences)
- normative knowledge (what society are we aiming for?)

In particular, Normative knowledge:

As we have already seen, it has to be driven by both general ethical theory and law regulations. We saw a lot of general ethical theories (AI for people, etc...) and now we get into a more precise aspect, which plays an important role: human values and rights.

HUMAN RIGHTS:



They are a big component but not sufficient to design the ethical lines to follow in the ICT revolution. NB: same rights can be seen in different ways depending on the culture (ex: privacy/reputation and freedom of speech in the EU vs in China) ⇒ human rights have an impact on the global perspective, like a framework within which the world tries to find an agreement.

Definition of HR (by Amartya Sen, indian philosopher):

These are primarily ethical demands, that should not be “juridically incarcerated”, and concern many kinds of *freedoms* (opportunities, like liberty and social rights), of which we need to satisfy some “threshold condition” of special importance and social influenceability.

HR are ethical, political (provided by society) and legal (granted by society) rights.

ICT and human rights: relationship

- interfere with HR
- contribute to protect/implement HR
- provide existence of new ones (ex: internet access)

Human rights that are important to look at in our field:

- Freedom and Dignity
 - *All human beings are born free and equal in dignity and rights.*
 - AI and technologies increase or decrease freedom and dignity?
 - increase ⇒ enables, gives new possibility (examples: in education and formation; socials and emails enables communications) even thinking to the definition of freedom that involves the “non-domination” aspect
 - decrease ⇒ more controllable (like in China); may be subject to systems as mere means losing dignity; AI doing people’s job; DM by machines with no possibility to know motivations (like not getting a bank loan); etc.
- 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.*
 - Technology gives opportunities in this sense but there is also the risk of magnifying differences already present in the real world... More if in the absence of adequate remedies
- 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 attack.*
 - AI and technologies enlarge and enhance the amount of collectable data, which is fundamental to gain new knowledge and possibilities, but it also brings risks in relation to privacy.
 - As a consequence, data has become a valuable resource, therefore it is object to norms, etc.
 - Corollary rights: EU “right to identity”, EU “right to erasure” or “right to be forgotten”, etc.
- Right to Life, Liberty and Security
 - *Everyone has the right to life, liberty and security (related to physical integrity) of a person.*
 - AI and technologies naturally affect this right... example: AV, intelligent weapons, etc. But also Medical Instruments, and others... [my example: many applications may help in feeling safe, for example, there is an app where women can trace themselves while moving on their own and report risky streets and so on <https://getbsafe.com/company/>]
- Right to Property
 - Everyone has the right to own property alone as well as in association with others.
 - Right to portability
- 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.*
 - Can AI interfere? yes, for example, it enables people in doing it + it can also enable governments or others in surveillance...

- 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.*
 - AI can hopefully effect this in accelerating, but: careful attention to the corruptancy of the systems, even the AI ones... Example: statistical interpretations (by AI systems) should not affect the interpretation of any person's single, specific case (⇒ COMPAS).
- 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.*
 - It may affect, for example with Automatic DM ... pay attention
- Presumption of innocence
 - *Everyone charged with a penal offence has the right to be presumed innocent until proven guilty according to law in a public trial at which he has had all the guarantees necessary for his defence.*
 - Is it approvable to intervene before the crimes are performed?
 - AI nowadays makes predictions about where crimes are more likely to happen... AI nowadays makes predictions about where home-violence could happen... should the systems intervene, getting somehow against this right?
- 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.*
 - Technology may give new opportunities (socials: people communicating; wikipedia; etc.), but this expanding liberty may also be better controlled thank to technologies (ex: China)
- 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.*
 - Connection between Political Rights and AI...
 - Driving idea: Freedom of Ideas on the internet like Freedom of Goods in economics
 - In some cases (ex: Cambridge Analytics), ICT and AI were used not to promote people's rational deliberation, but instead to manipulate and influence them... another example: big debates over advertising and so on on social media (for example reddit is not accepting paid political advertisement anymore, while facebook and others do...)
- 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/her personality.*
 - AI could reduce the cost of the management of social services and contribute to an effective, not invasive social security

- In general, society could become more efficient and productive if guided by a tool such as AI
- Right to Work
 - *Everyone has the right to work, to free choice of employment, to just and favourable conditions of work and protection against unemployment.*
 - could AI affect this right? Of course... With a negative impact due to redundancy (taxi drivers and self-driving cars) + with positive impact creating new opportunities (ex: Amazon, 840000 employees)
 - Again, in relation to Work, even the Dignity of workers may be in danger with AI surveillance (ex: Deliveroo). But AI could also reduce dangers in activities (getting machines to do them) or preventing risky situations (improving work conditions)
- 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 and ICT can contribute by increasing social productivity (ex: granting an adequate standard of living to everyone, etc.)
- Right to Education
 - *Everyone has the right to education. Education shall be free, at least in the elementary and fundamental stages.*
 - ICT and AI highly contribute to this right, think of covid pandemic and online lessons... or to information access, tools and so on.
- 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.*
 - ICT facilitates access to intellectual and artistic work as well as the creation of new contents
 - AI can play a role in this by providing new ways of expressing artistic ideas but also exploring science.

Conclusions

Human rights, as we have the ICT revolution are

- A precious heritage to protect, but also
- blueprints for a human centred ICT, and in particular human centred AI.

Human rights do not exhaust the planning for the future but they are crucial in this design, in addition they are less controversial than broad ethical theories.

LOGIC PROGRAMMING ARGUMENTATION FOR EXPLAINABLE AND ETHICAL AI (Calegari)

This section is about how to program and design ethical behaviour from a computer engineering perspective, in particular design and implement it through declarative and logic-based approaches.

Context

AI systems developed to be involved in a wide range of fields, where more complex issues are in increased demand of proper consideration, in particular when the agents face situations involving choices on moral or ethical dimensions and issues of responsibility

So... important aspects are INDIVIDUAL cognition, deliberation, and behavior

- + COLLECTIVE morals, and how they emerged

⇒ design a model of knowledge addressing also morality issues

Agents in LP programs

- agents = **autonomous** computational, programmed entities
 - genus = agents that are computational entities
 - differentia = agents that 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 only be reactive, intelligent, mobile
- they live within Multi-Agent Systems (MAS), and interact with other agents through communication actions, and with the environment with pragmatic actions

Motivation: Why Logic-based Approach?

Main purpose of AI applications: design and incapsulate intelligence

LP is an alternative way of delivering symbolic intelligence, complementary to sub-symbolic approaches.

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

- moral permissibility, taking into account the doctrines of double effect and triple effect, and Scanlonian contractualism
- the dual process model that stresses the interaction between deliberative and reactive processes in delivering moral decisions
- the role of counterfactual thinking in moral reasoning

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

Technically, we will see:

- 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?” ⇒ 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 trusted and safe intelligent systems that can be critical in the design of ubiquitous intelligence (both in terms of transparency and in terms of ethics):

- PROVABILITY - it can provide proof to its response for a well-founded-semantic model
- EXPLAINABILITY - it gives formal methods for argumentation and justification
- EXPRESSIVITY + SITUATEDNESS - it permits extensions, explicitation of exceptions and assumption and captures specificities of contexts
- HYBRIDIZATION - it allows diversity integration

NB: not an agent-programming (like json) but LP because of logical inference, necessary for reasoning, deliberation, etc. with to give to agent-operations. Logic is good to represent goals, plans and so on. In addition, LP permits to build cognitional artifacts (MAS theory).

Preliminaries

⇒ PROLOG (Horn clauses, depth-first search strategy, automatic backtracking)

NB: since depth-first strategy, order of the clauses, etc. is important from both computational and ethical-objective perspective.

Abduction extension

Step of adopting an hypothesis as being suggested by the facts

Abductive program format: <P, AB, IC>

P = logic program

example: *Grass is wet if it rained.*
 Grass is wet if the sprinkler was on.
 The sun was shining.

AB = set of predicates names (abducible predicates)

example: *Grass is wet.*

IC = set of first-order classical formulae

example: *false if it rained and the sun was shining*

⇒ abduction: *The sprinkler was on.*

Abstract Argumentation

An argumentation system consists of a couple (A,R) where A is a set of elements and R is a binary relation representing attack relation between arguments:

An LP approach to Ethics

Abduction:

It enables the generation of plausible scenarios to be generated under certain conditions, and enables hypothetical reasoning, including the consideration of counterfactual scenarios about the past.

Counterfactual reasoning suggests thoughts about what might have been, what might have happened if any event had been different in the past.

It provides hints about the future by allowing for the comparison of different alternatives inferred from the changes in the past.

It also provides justification of why different alternatives would have been worse or not better, and integrity constraints which exclude abducibles that have been ruled out a priori. A posteriori preferences are appropriate for capturing utilitarian judgment that favors welfare-maximizing behaviors. It combines a priori integrity constraints and a posteriori preferences, resulting in a model which reflects the dual-process of intuition and reflection. A priori integrity constraints are mechanisms to generate immediate responses in deontological (a priori) judgement.

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 (PLP):

It enriches symbolic reasoning with degrees of uncertainty.

It allows abduction to take scenario uncertainty measures into account.

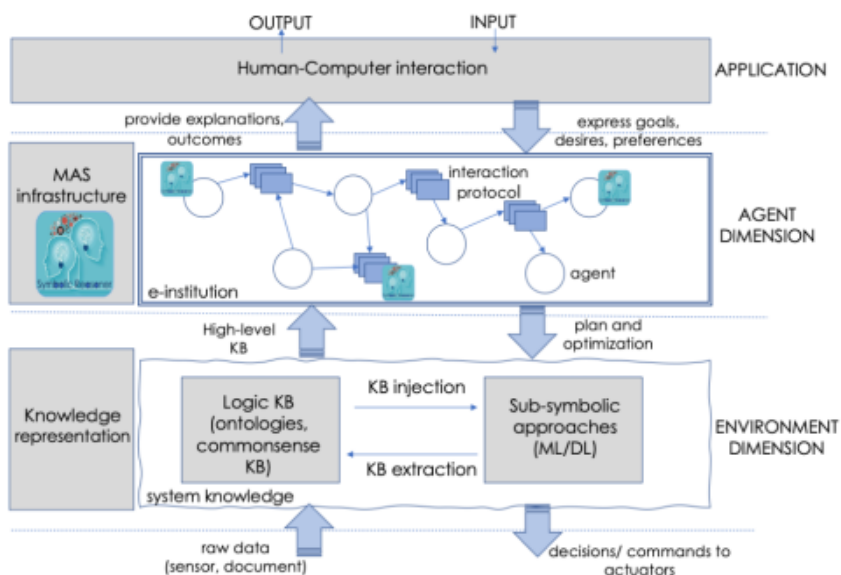
It accounts for diverse types of uncertainty, in particular uncertainty on the credibility of the premises, uncertainty about which arguments to consider, and uncertainty on the acceptance status of arguments or statements.

One of the key factors that allow a system to fully meet, managing to formulate well-founded reasoning on which scenario to prefer and which suggestions to provide as outcomes.

Argumentation:

It enables system actors to talk and discuss in order to explain and justify judgments and choices, and reach agreements. Despite the long history of research in argumentation and the many fundamental results achieved, much effort is still needed to effectively exploit argumentation in a distributed and open environment.

Possible Architecture



NB: knowledge representation is a combination of both symbolic and subsymbolic techniques!

Calegari suggested **tuProlog** because it is a java-based platform which allows exploitation of subsymbolic techniques as well.

12/04/2021 - MODELLING NORMS (Contissa) + A Model for Rules (Contissa, Galileo Sartor)

Knowledge representation in particular in the legal domain, so concerning legal reasoning and representation of legal norms or norms in general, namely ethical norms.

MODELLING NORMS (Contissa)

Historical insight

At the time of birth of AI, already with McCarthy and the others, the Law/Legal domain appeared to be one of the first applicative contexts for AI (along with medicine).

So there were created man-made models of the law, on the idea of “computable law”.

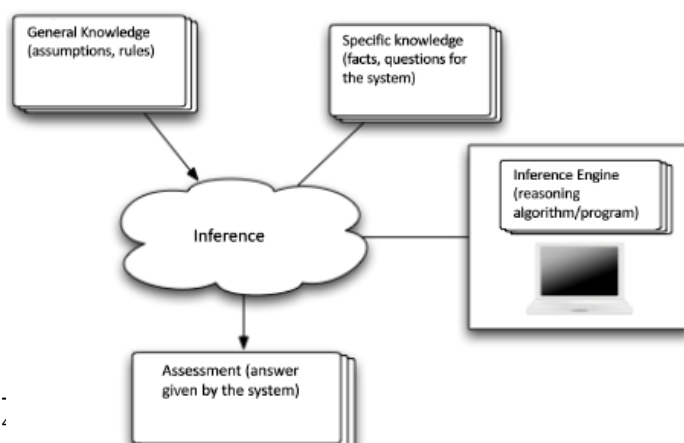
- HOW? Modeling/formalisation of the law...
 - Input: sources, cases (jurisprudence), concepts, doctrines ⇒ of course adequately translated in some way understandable by the machine
 - Output: computable models (knowledge base)
 - Process: logic programming/knowledge representation
- TO WHAT AIM? ... in order to use the previous output (computable models) as a new input to get applicable provisions:
 - Input: computable models of the law
 - Output: Answers, legal qualifications, support to decision-making
 - Process: Forward and backward rule chaining, deduction, defeasible reasoning, etc.

So the first process, namely logic programming/knowledge representation consists in the application of logic and ontology⁴ to the task of constructing computable models for some domain ⇒ Logic: provides the formal structure and rules of inference.

Ontology: defines the kinds of things that exist in the application domain, and their interrelationship.

Computable models: implement logic and ontology into computer systems and applications **which are computable**.

Technically speaking this was generally achieved by using symbolic representation, like with Declarative programming languages (ex: Prolog), consisting of logical statements which express the knowledge about the domain in terms of known facts and relationships. This kind of program is executed by searching for proofs of the statements.



This approach is also used to develop one kind of symbolic AI application in particular: Rule-based systems and specifically to our interest ⇒ Legal Rule-based systems

Legal applications:

- By the 1980s, a number of researchers had implemented working systems based on manually created logical representations of rules e.g., Sergot et al. (1986) (British Nationality Act)
- Nowadays Rule base systems are used in the legal domain for legal analysis and automated legal assessment, and this is due not only to technological reasons but also to the way laws are “produced”, and in line with this aspect, they are particularly successful in the anglosaxon contexts.
 - ALTHOUGH WE CANNOT SUBSTITUTE LAWYERS AND JUDGES WITH LOGIC FORMALISMS AND COMPUTABLE MODELS, HOWEVER THERE ARE MANY CONTEXT IN WHICH THE SIMPLE APPLICATION OF THE RULES IS THE CORE CONTENT OF THE LAW, for example many applications in public administration, like taxes, welfare, one-stop shop for enterprises, online legal proceedings, etc., and in business application, ex: business rules.

Legal domain and Rule-based systems

Why have law systems been considered an example of a rule-based systems scheme?

Because we can think of legal rules as a set of conditional statements (not all of it of course), consequently much of the legal reasoning is translatable to logical propositions.

Examples:

- PENAL LAW: *if a person Y commits the crime X, then Y shall be punished with sanction Z*
- CIVIL LAW: *If X buys the good Z from Y, then X shall pay to Y the price of Z.*
- CIVIL RIGHTS: *If X was born in the territory of State S, then X is a citizen of S.*

NB: in this kind of expressions (but could be translated), **not using quantifiers... ⇒ pseudo-predicative representation**

So the form of predicates is

- Premises...
 - (RULE): *If X buys Z from Y, then X shall pay to Y the price of Z.*
 - (FACT): *Mary buys a car from John.*
- ... plus Conclusion:
 - (LEGAL EFFECT): *Mary shall pay to John the price of the car.*

Legal knowledge representation: issues and challenges

- **Ambiguity:** legal rules may be ambiguous (see later on).
- **Vagueness:** it concerns meaning of the rules (so also related to ambiguity), saying that the representation might leave out interpretation uncertainties.
- **Rigidity:** logic representation is rigid, by means it crystallizes considered situations in a symbolic structure, that makes it difficult to give a characteristic of laws: that it always provides a solution even if it has to stretch what is available in order to fit into the rules.

- How to **represent deontic positions**: the law is not concerned with observing the reality as it is but how it ought to be (so deontic logic⁵).
- How to **enable temporal reasoning** (law as a dynamic system): in law systems, there is the necessity to make temporal reasoning in different perspectives. Example: “*If X earns money Y during the current year, X will owe Z in taxes the following year*”; in this case there are internal times - 1) the time of current year, 2) the time of following year - and external times - 3) the time this norm was not active, 4) the time this norm was enforced, 5) maybe the time this norm has been abrogated/amended, etc.
- How to deal with **conflicting legal rules** and/or rules that can be excluded from being applicable by other rules: many times, legal systems allow to correctly deduct conflicting legal effects starting from the same fact and applying different rules ⇒ legal systems are not coherent and sound... For this reason there are some ploys, some meta-rules (⇒ **meta-reasoning**), like “*lex superior derogat inferiori*” or “*lex specialis derogat generali*”, that define which rule has to be followed in presence of conflicting legal effects.
- How to **manage reification**, whenever rules representing legal norms need to be treated as objects with properties by other rules: not only I can have conflicts between rules, but also there might be rules that can be excluded from being applicable by other rules (defeasibility of a rule), so there is need also for **defeasible reasoning**⁶. So... temporal reasoning + meta-reasoning + defeasible reasoning ⇒ need to manage reification⁷.
- How to **maintain isomorphism**⁸ between source, text and representation: this is a practical problem and a legal requirement, because law official source (like Gazzetta Ufficiale in Italy) - although maybe ambiguous, vague and so on - is binding for what the law states, so isomorphism to these sources must be preserved, despite the contradictions, changes over time, etc., that these sources may report.
- Other more practical issues: knowledge elicitation/representation/update bottleneck...

⁵ https://en.wikipedia.org/wiki/Deontic_logic Deontic logic is the field of philosophical logic that is concerned with obligation, permission, and related concepts. [...] Typically, a deontic logic uses OA to mean “*it is obligatory that A*” (or “*it ought to be (the case) that A*”), and PA to mean “*it is permitted (or permissible) that A*”.

A deontic expression indicates:

- the state of the world not meeting some standard or ideal ⇒ how the world ought to be according to certain norms, expectations, speaker desire, etc.
- and some action that would change the world so that it becomes closer to the standard or ideal ⇒ some obligation, permission, or related concepts.

⁶ https://en.wikipedia.org/wiki/Defeasible_reasoning a kind of reasoning that is rationally convincing, though not deductively valid. [...] A non-demonstrative reasoning, where the reasoning does not produce a full, complete, or final demonstration of a claim.

⁷ in knowledge representation, the process of turning a predicate into an object + in natural language processing, the process of transforming a natural language statement so that actions and events in it become quantifiable variables.

⁸ <https://en.wikipedia.org/wiki/Isomorphism> structure-preserving mapping

EXAMPLES: 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"*

Pseudo-logic Translation:

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

Problem: ambiguous meaning of the previous

- *IF {(a AND b) OR (c AND d)} THEN e?*
- *IF {(a OR c) AND (b AND d)} THEN e?*
- *IF {(a OR c) AND (b OR d)} THEN e?*

The interpretation problem is in "such system"!!

EXAMPLES: Vagueness

"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, Chapter VI)

"No vehicles allowed in the park"

- Core meaning: cars, motorbikes, coaches, etc.
- Penumbra: bikes, skateboards, horses, trolleys?
- Additional Penumbra: what about emergency vehicles in action?

(We had a look to "The British Nationality Act as a Logic Program" paper
⇒ download the pdf on Virtuale, it also has Prof's highlighting)

A modern systems: Oracle Policy Automation (OPA⁹)

This is a system that was originally developed by RuleBurst then acquired by Oracle. It is a 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. Then there is a linguistic component (parser) that analyses the syntactic structure of phrases in order to identify their logical components. The rules are then translated into an XML-based format, used by the Inference Engine. The linguistic component also automatically prepares questions and explanations for the user interface.

A Model for Rules ⇒ SWI-Prolog (Contissa, Galileo Sartor)

- SWI-Prolog: a promising approach for our purposes
- They modelled the *Brussels*, *Rome I* and *Rome II* regulations
- In the future, they will add national or international norms, and see how they interact with the existing rules
- A link between the reasoning component and the database of cases and legislation must be defined and implemented

⁹ <https://www.oracle.com/middleware/technologies/oracle-policy-automation-downloads.html>

19/04/2021 - PRIVACY AND DATA PROTECTION

General Data Protection Regulation

HUMAN RIGHT: 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 attack.”

AI and technologies enlarge and enhance the amount of collectable data, which is fundamental to gain new knowledge and possibilities, but it also brings risks in relation to privacy. As a consequence, data has become a valuable resource, therefore it is subject to norms, etc...

The European Union, in 2018, approved the GDPR. It consists of 99 Articles and it states sanctions depending on the gravity of infringements.

The GDPR norms apply to any organization operating inside the EU and to any organization outside of the EU which offer goods or services to customers in the EU.

GDPR appeared necessary because protection of natural persons with regard to the processing of their personal data was needed in order to guarantee the Right to Privacy and Data Protection, which is a fundamental right and freedom of natural persons.

Article 4 - Subjects of the GDPR

- Data subject: identified or identifiable natural person.
- Data controller: natural or legal person, agency or other body which determines the purposes and means of the processing of personal data.
- Data processor: natural or legal person, agency or other body which processes personal data on behalf of the controller.

Article 5 - Personal data

Personal data (information related to an identifiable natural person) shall be:

- processed lawfully and in a transparent manner
- collected for specified purposes and not further processed in a manner that is incompatible with those purposes
- limited to what it is necessary in relation to the purposes
- accurate and up to date
- kept in a form which permits identification of data subjects for no longer than is necessary for the purposes
- processed in a secure way to avoid unauthorized or unlawful access

The controller shall be able to demonstrate compliance with the above requirements.

Article 6 - Lawfulness

Processing shall be lawful only if one of the following applies:

- the data subject has given consent to the processing
- the processing is needed for the performance of the contract with the data subject
- processing is necessary for the performance of a task carried out in the public interest

Article 7 - Categories of personal data

There are special norms related to the processing of special categories of personal data (ethnicity, political opinion, biometric data, health data, ...)

Article 13 - Data Controller obligations

When personal data are collected from the data subject, the controller shall provide the following information:

- who is the controller and its contacts (identity)
- contact details to the data protection officer
- the purposes of the processing
- the period for which the personal data will be stored

Article 17 - Right to be forgotten

Right to be forgotten: the data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her.

Article 20 - Right to data portability

Right to data portability: the data subject shall have the right to receive the personal data concerning him or her.

Article 22 - Data Subject rights

The data subject shall have the right not to be subject to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her.

Article 25 - Data protection by design and by default

Data protection by design and by default: the controller shall implement appropriate measures to implement data-protection principles, such as data minimization. In addition, only personal data needed for each specific purpose shall be processed.

Article 82 - Infringements and compensations

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

Article 83 - Effective, proportional and dissuasive Infringements

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 (related to technical and organizational measures)
- the previous violation by the data controller
- cooperation with supervisory authority
- affected categories of personal data

GDPR and AI

The GDPR does not speak about AI however it contains many relevant provisions and it can be interpreted and applied so as to address risks and allow for all kinds of beneficial uses of AI.

But ask yourself... Do we really need Automated price discrimination or Face recognition or Psychography/Emotion detection or Targeted political advertising??

- GDPR is the vanguard of the regulation of AI; other domains of the law are must follow
- AI is key to the future of Europe, to its economic success and to the well-being of its citizen
- GDPR contributes to ensure the beneficial deployment of AI while preserving and enhancing human values
- The GDPR needs to be complemented by detailed technical regulations as well as high level soft and hard provisions based on a broad political and social debate.

26/04/2021 - FAIRNESS IN AUTOMATED DECISION MAKING + COMPAS (Lagioia)

1. AI in decision making concerning individuals
 - a. possible causes of unfairness
2. principle of fairness and its substantive dimension
3. AI unfairness
 - a. the COMPAS predictive system and the Loomis case
 - b. a toy example and the criteria for assessing fairness

1. AI in decision making concerning individuals

Fairness and discrimination:

AI+big data ⇒ automated decision making even for complex decisions, according to multiple factors and non-predefined criteria

WIDE DEBATE due to both perspectives and risks of algorithmic assessments and the impact of automated decision making over individuals

WHY YES automated decision making? not only cheaper but also possibly more precise (humans' inability to process statistical data) and impartial (humans' typical prejudice, overconfidence, loss aversion, bias representativeness, etc)

SOMEONE UNDERSCORED POSSIBILITY OF ALGORITHMS MISTAKING OF DISCRIMINATING ⇒ BUT NO, there are very few cases of algorithms engaging in explicit unlawful discriminations and in fact systems' usual standards are considered to perform better than human experts. The general biggest risk is systems being disproportionately affecting certain groups with no acceptable rationale ⇒ generally due to systems reproducing weakness of human judgement (including errors and prejudice) due to their learning techniques like supervised learning, which is based over data representing past human biased behaviour:

- biased rules
- fair rules but biased training set (example: AMAZON HIRING SYSTEM - skilled candidates' but facilitating white and male due to historical hirings)
- biased statistical composition of datasets' population

How to challenge this kind of unfairness? Well, this is, in fact, considered to be very difficult due to the risk of machines rejecting humans' interference because it would raise additional cost and uncertainty: statistical correlation (the basis of systems' behaviour) justifies exceptions.

AI-supporting experts' idea is to regulate systems with "putting in place the right safeguards" to manifest the potentiality of systems of being the positive force for equity.

So... putting in place the right safeguards ⇒ integrating human and automated judgements! To move in this direction, the main development focus has to be to make it easy to examine and interrogate systems over their decision process. By the way, machines are at least controllable, "measurable" and can be engineered with respect to humans...

2. Principle of Fairness and its Substantive Dimension

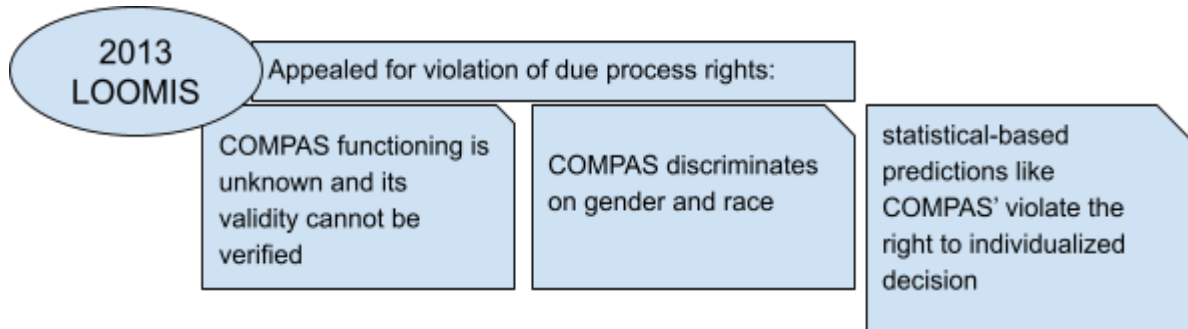
Equal and just distribution of benefits and costs + individuals and groups free from unfair bias, discrimination and stigmatisation + AI decision making

So best approach so far: fairness ⇔ transparency + explainability

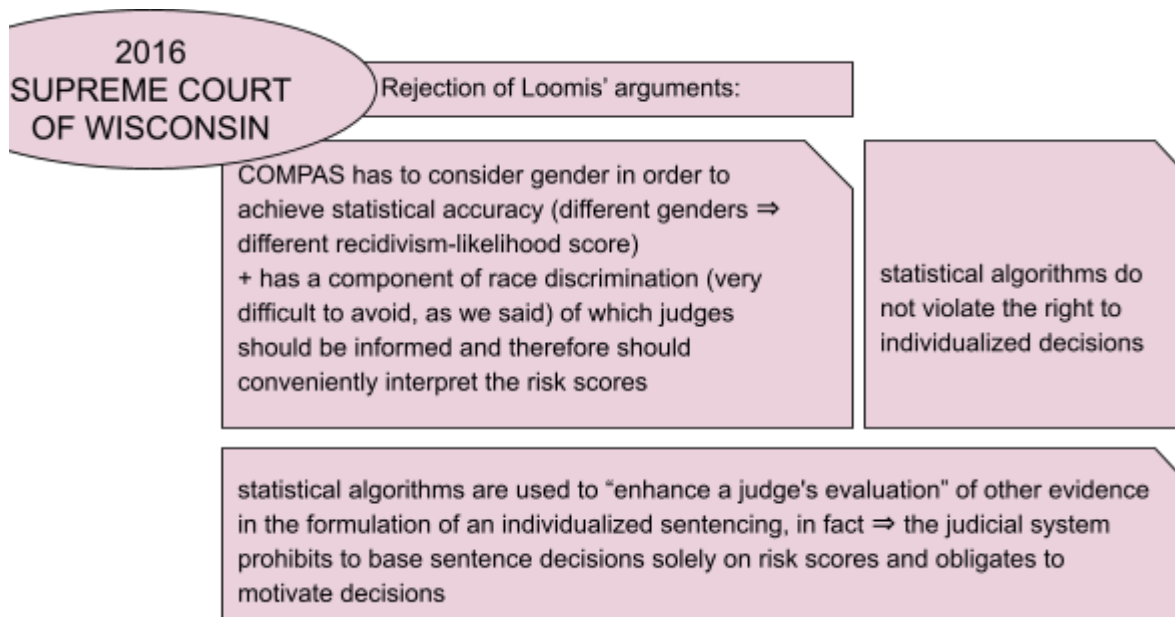
3. AI unfairness

COMPAS SYSTEM - It is a risk assessment tool used by American judges to state the probability of offenders to be recidive and to decide the most appropriate correctional treatment. It is based on statistical algorithms which explore, learn and predict on the basis of a multiple choice test, static risk variables (prior criminal history, education, etc.) and dynamic risk variables (drug abuse, employment status, social integration, etc.) ⇒ the prediction says if the defendant has a low, medium or high risk of being recidive.

In 2013, E. Loomis stole a vehicle and fled the police. He was classified as high risk and sentenced to 6 years imprisonment also after COMPAS risk assessment:



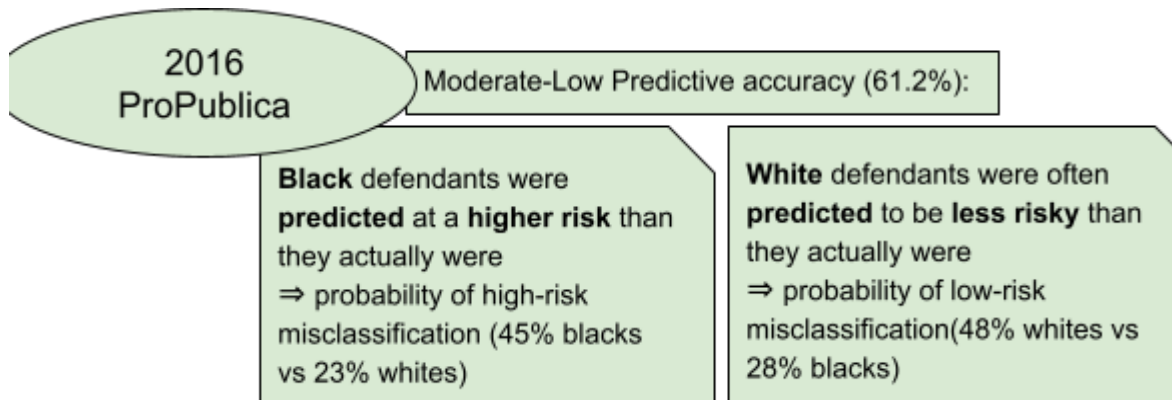
In 2016, the Supreme Court of Wisconsin responded to his appeal:



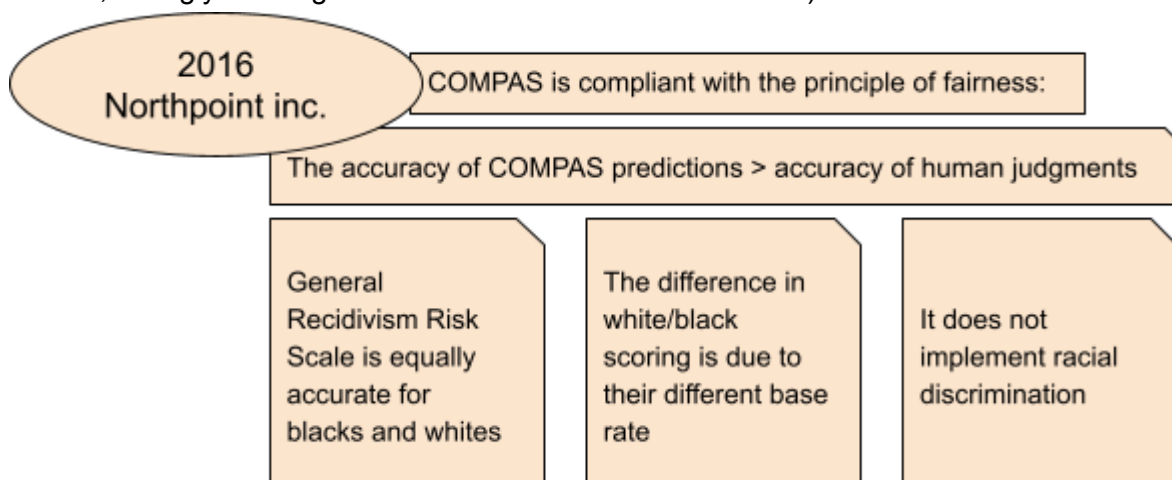
In 2016 ProPublica¹⁰ published a study:

- evaluating COMPAS accuracy and fairness
- on the basis of a sample of 12000ish defendants
- by comparing COMPAS' predicted recidivism rates in 2013/2014 and the rate that actually occurred in 2016 for those defendants

¹⁰ <https://www.propublica.org/> <https://it.wikipedia.org/wiki/ProPublica> It is a nonprofit organization-newsroom that aims to produce investigative journalism in the public interest. Five-times Pulitzer Prize winner.



And in the same year, Northpoint¹¹ (contributors to COMPAS), responded to ProPublica, saying they made several statistical and technical errors (like misspecifying regression models, wrongly defining classification terms and measures...):



NB: the base rate is like an intrinsic starting point which is static and computed on the basis of factors like poverty, unfavourable social conditions, etc.

OUR EXPERIMENT: SAPMOC (??)

Real Outcomes			
	Recidivism	No Recidivism	Total
Previous Offence	750	250	1000
No Previous Offence	250	750	1000

SAPMOC Predictions			
	Recidivism	No Recidivism	Total
Previous Offence	1000	0	1000
No Previous Offence	0	1000	1000

SAPMOC system main aspects:

- Statistical Parity: NO
 - Each group has equal proportion of positives and negatives predictions
- Equality of Opportunity: NO

¹¹ <https://northpoint-inc.com/> https://en.wikipedia.org/wiki/NorthPoint_Communications a competitive local exchange carrier focused on data transmission via digital subscriber lines.

- The members of each group, which share the same features, should be treated equally in equal proportion
- Calibration: OK
 - The proportion of correct predictions should be equal within each group and with regard to each class
- Conditional Use Error: OK
 - The proportion between FP (FN) and the total amount of positive (negatives) predictions should be equal for the 2 groups.
- Treatment Equality: NO
 - The ratio between errors in positive and negative predictions should be equal in all groups

What we look for:

- Equal accuracy within groups
- Different base rate explains the violation of statistical parity, treatment equality, and equality of opportunities
- Violation of fairness criteria does not necessarily lead to unfairness
- Shall we impose statistical parity? (Lower accuracy + higher false rate + discrimination against individuals) ⇒ ex: quote rosa
- Individuals fairness vs group fairness

How to achieve it:

- Unpacking the decision
 - Unfairness in prediction (prohibited features, biased data set, biased proxy, etc.)
 - Unfairness in classification (threshold – affirmative actions)
 - Unfairness in decision (right/values optimization)
- Predictive systems as instruments to understand the reality

FUTURE: AI is too 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

03/05/2021 - AUTONOMOUS VEHICLES (Fossa)

Autonomous driving: ethical and social issues

Introduce autonomous driving and its ethical significance

There are many different levels of autonomous driving:

	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	
What do these features do?	These are driver support features			These are automated driving features		
	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	<ul style="list-style-type: none"> • automatic emergency braking • blind spot warning • lane departure warning 	<ul style="list-style-type: none"> • lane centering OR • adaptive cruise control 	<ul style="list-style-type: none"> • lane centering AND • adaptive cruise control at the same time 	<ul style="list-style-type: none"> • traffic jam chauffeur 	<ul style="list-style-type: none"> • local driverless taxi • pedals/steering wheel may or may not be installed 	<ul style="list-style-type: none"> • same as level 4, but feature can drive everywhere in all conditions

Nowadays implementations are of level2 or at most lever3 autonomous-ness.

The main problem is integration of autonomous capabilities.

Provide an overview of the Ethics of Autonomous Vehicles

- **What?** Technical goal and issues: Level 5 Autonomy

But, once we go into the real issues, we understand that the technical aspects are not so central, they are the “easily solvable” ones...

- **Why?**
 - Autonomy and Freedom ⇒ opportunity to do else in the time we usually spend driving
 - Safety ⇒ get rid of the human error (90% of the accidents)
 - Sustainability
 - Inclusiveness ⇒ design AV so that people who are generally excluded from car-experience get now included, for example people with disabilities
- **Who?**
 - Everybody?
 - Only those who can afford them?
 - Only those who can take control, if needed?
- **How?**
 - Social Trust
 - Responsibility

- Rights
- By Law / Discretionary
- New Possibilities
- **Where?**
 - Everywhere
 - Highways
 - Parking lots
 - National / International

Ethics of AV:

1. unavoidable collisions
2. privacy & security
3. responsibility allocation
4. sustainability
5. personal freedom and the social good

How to face all of these? DESIGN PROCESS INVOLVING ETHICS, POLITICS and NORMATIVES.

Regulation and Policy- Ethics of Connected and Automated Vehicles:

- Released on Sept. 17th, 2020
- Authored by an Independent Expert Group – 14 members, mostly academic (philosophy, law, engineering)
- Establishes a baseline for future European policy on connected and automated vehicles
- Ethical & social issues are widely accounted for

Debate of some specific ethical problems:

1. **unavoidable collisions** ⇒ Q: *how should the system handle morally laden situations?* – i.e., situations where harm is unavoidable but can be distributed in different ways ⇒ A: *Accident-algorithms*, this means that we have the possibility to design and decide how to cope and what to do during an accident, the possibility to take action... but what shall be done? This is also an ethical problem:



Many issues:

- Which value/ethical theory to implement?
- How could we do that?
- Who gets to decide?
- How should this choice be made?
- What about personal autonomy?
- What about the rights of bystanders?

- Many problems:
- Sometimes a bit detached from vehicle dynamics
- Sometimes a bit detached from available/foreseeable tech
- Sometimes a bit detached from realistic scenarios

⇒ Still the problem remains!

2. **privacy & security**

- a. For autonomous vehicles to function properly, a huge quantity of data must be collected, shared, and stored... ⇒ Privacy protection throughout the entire infrastructure + need for informed consent
- b. Autonomous vehicles pose risks proper of both usual vehicles and information systems
- c. double challenge:
 - i. “Mechanical” vehicle safety standards
 - ii. Digital infrastructure liabilities: external attacks (security), software issues (robustness), data thefts and leaks (...)

⇒ problems like AV fooling with phantom images, hackers could take control over the AV’s sensors, hacking street signs with stickers and so on could confuse the AV, etc.

3. **responsibility allocation** ⇒ Q: *Who is to be held responsible for harm caused by accidents where autonomous vehicles are involved?* A: *NOT the systems themselves...* then who? passengers? owners? designers/developers? producers? nobody, just the insurance system?

- a. Meaningful Human Control Approach:
 - i. Autonomous vehicles 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!

4. **sustainability**

- a. Environmental Impact:
 - i. More or less vehicles in use?
 - ii. Materials are reusable or recyclable?
 - iii. Energy consumption (data centres)?

⇒ it is believed that self-driving cars would be more sustainable
- b. Social Impact:
 - i. Disabilities and minorities
 - ii. More or less traffic?

⇒ it is believed that self-driving cars would be more equal
- c. Economic Impact:
 - i. Job losses / New jobs?
 - ii. Who can afford AVs? Is this fair?

⇒ if they are safer, they should not be more expensive than regular cars, in order to widely spread

5. **personal freedom and the social good**

- a. Value conflicts:
 - i. Safety vs Pleasure
 - ii. Personal privacy vs System efficiency
 - iii. Moral autonomy vs Human error
 - iv. Passenger protection vs Bystanders’ rights
- b. Should human driving be outlawed?
 - i. Yes: minimize road casualties

- ii. Yes: maximise traffic efficiency
- iii. No: individual freedom
- iv. No: discrimination

Discuss your impressions, questions, doubts, perplexities and suspicions

He asked to answer his last question, someone said yes and maybe create specific circuits where passionate drivers can still drive... he said this is something that has actually been proposed!

Observation: elder people might not be ready to cooperate and trust AV... There is in fact the need to consider different social groups!

Why not to have both? AV and human-driven vehicles? not recommended because would generate accidents: humans are not predictable so there would not be the AV's full advantage

Observation: with level5 AV we could get rid of traffic lights, plus maximal speed could be increased (even if speed is not a common concern) thanks to shared data about cars' position, etc., and there would not be traffic basically for the same reason.

Why not implement an AV that is drivable by humans but decide autonomously for those aspects that are related to accidents? This is probably because the automated features in cars come from the opposite direction: it should offer assistance, it is not the opposite with cars deciding and humans giving assistance...

10/05/2021 - CLAUDETTE SYSTEM (Lagioia)

Context

Recently, the popular perception of AI is that of something at the service of businesses, currently affecting consumers: privacy, autonomy, economic interests, behaviour, etc. That - as we saw in loads of lectures... - does not have to be the case!

So... How to empower consumers?

- Protection against unwanted monitoring (GDPR)
- Support in detecting unfair use of AI
- Control commercial practice fairness

This is where the idea of CLAUDETTE was born... It is a system developed by some researchers across the EU (like Sartor and Lagioia) and in particular with our DISI department (also Torroni works at it!).

“CLAUDETTE” stands for “clause detector”, this is because it is a ML system that automatically **detects clauses** that might be potentially unfair in Terms of Services and Privacy Policies!, which is very important because:

- generally consumers agree but don't read
- NGOs (Non-Governmental Organizations) have competence to control but lack resources
- Business keeps using unlawful clauses

CLAUDETTE is now available as an online server: <http://claudette.eui.eu/>

Technological aspects

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

- a detection task: does a sentence contain a potentially unfair clause? Positive (if p unfair), Negative (otherwise)
- a sentence classification task: what is the category the unfair clause belongs to?

Implemented approaches:

- + Bag of Words (BoW): build to leverage the lexical information in sentences
- + Tree kernels: structure of sentences by describing the grammatical relations between sentence through a tree
- + Convolutional Neural Networks, SVM, etc.

Leave-One-Out procedure: each document in turn, is used as a test set, leaving the remaining documents for training set (4/5) and validation set (1/5) for model selection.

Terms of Service (ToS)

As we said, CLAUDETTE is a ML system therefore it needed a training set to learn how to recognise and label correctly the clauses.

According to CLAUDETTE group of research, clauses within a Contract are divided in:

1. clearly fair
2. potentially unfair
3. clearly unfair

The definition of an unfair clause is in EU's 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.”

⇒ there are some types of clauses that traders are prohibited from using in the contracts.

Each clause of a contract gets labelled by CLAUDETTE system with xml tags representing the category over which the clause appears to be unfair and the number corresponding to the level of unfairness (1 ⇒ clearly fair, 2 ⇒ potentially unfair, 3 ⇒ clearly unfair).

In particular, clauses can be considered fair/unfair according to one aspect but to another... This is why CLAUDETTE team decided to state 8 different categories of unfairness:

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	<ld>	296	49
Unilateral termination	<ter>	236	48
Contract by using	<use>	117	48

Examples for CONTRACT BY USE category:

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: 2 ⇒ potentially unfair

A potentially unfair consent by using clause (Facebook):

```
<use2>By using or accessing the Facebook Services, you agree to this Statement, as updated from time to time in accordance with Section 13 below.</use2>
```

Examples for JURISDICTION category:

If giving consumers a right to bring disputes in their place of residence: 1 ⇒ clearly fair

If stating that any judicial proceeding takes a residence away (i.e. in a different city, different country): 3 ⇒ clearly unfair

A clearly unfair jurisdiction clause (Dropbox):

```
<j3> You and Dropbox agree that any judicial proceeding to resolve claims relating to these Terms or the Services will be brought in the federal or state courts of San Francisco County, California, subject to the mandatory arbitration provisions below. Both you and Dropbox consent to venue and personal jurisdiction in such courts.</j3>
```

Examples for LIMITATION OF LIABILITY category:

If stating that the provider may be liable: 1 ⇒ 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”: 2 ⇒ potentially unfair

If stating that the provider will never be liable for physical injuries (health/life)// gross negligence// intentional damage: 3 ⇒ clearly unfair

A fair jurisdiction clause (World of Warcraft):

```
<lt;1>Blizzard Entertainment is liable in accordance with statutory law (i) in case of intentional breach, (ii) in case of gross negligence, (iii) for damages arising as result of any injury to life, limb or health or (iv) under any applicable product liability act.</1>
```

A potentially unfair jurisdiction clause (9gag):

```
<lt;2>You agree that neither 9GAG, Inc nor the Site will be liable in any event to you or any other party for any suspension, modification, discontinuance or lack of availability of the Site, the service, your Subscriber Content or other Content.</2>
```

A clearly unfair jurisdiction clause (Rovio):

```
<lt;3>In no event will Rovio, Rovio's affiliates, Rovio's licensors or channel partners be liable for special, incidental or consequential damages resulting from possession, access, use or malfunction of the Rovio services, including but not limited to, damages to property, loss of goodwill, computer failure or malfunction and, to the extent permitted by law, damages for personal injuries, property damage, lost profits or punitive damages from any causes of action arising out of or related to this EULA or the software, whether arising in tort (including negligence), contract, strict liability or otherwise and whether or not Rovio, Rovio's licensors or channel partners have been advised of the possibility of such damages.</3>
```

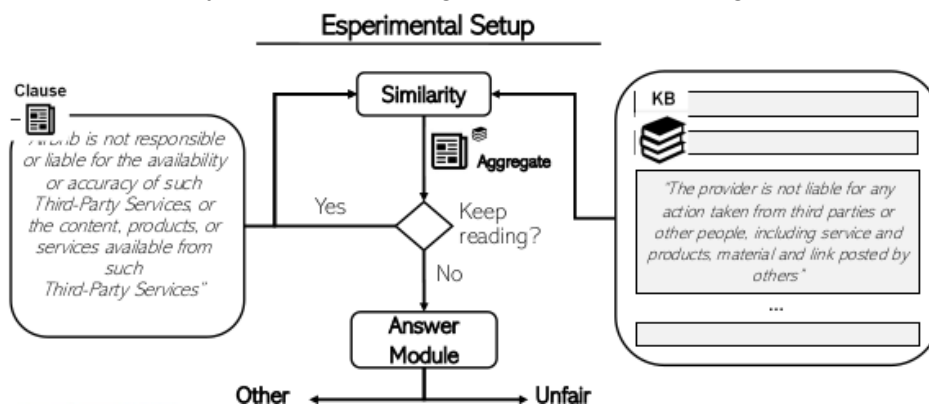
NB: Human Legal experts are able to recognize potentially unfair clauses thanks to their background knowledge of the domain ⇒ Able to explain their intuitions of unfairness, provide reasons why a clause is unfair (Legal Rationales), and use rationales to guide such intuitions ...How to simulate the same for a machine? Memory-Augmented Neural Networks

Memory-Augmented Neural Networks

- Process input and store the information in some memory
- Understand pieces of knowledge relevant to a given query
- Retrieve concepts (*legal rationale*) from memory, the Knowledge Base (KB)
- Combine memory and query to make a prediction

The prediction is made by comparing and stating the similarity between new clauses and the knowledge base of the model.

Each clause may be passed through multiple times as a guarantee (*enhanced input*).



CLAUDETTE and GDPR

GDPR is the golden standard of Lawfulness, Fairness and Transparency. To state that a policy is GDPR-compliant, it must satisfy 3 dimensions:

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

Type of required information	Symbol
Identity of the controller (controller's representative)	<id>
Contact details of the controller (controller's representative)	<contact>
Contact details of the data protection officer	<dpo>
Purposes of the processing	<purp>
Legal Basis for the processing	<basis>
Categories of personal data concerned	<cat>
Recipients or categories of recipients of the personal data	<recep>
Period for which the personal data will be stored, or the criteria used to determine that period	<ret>
Right to lodge a complaint with a supervisory authority	<complain>
...	<...>

Example for 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>

<dpo2> ⇒ “Contact details of the data protection officer” + potentially unfair

The clause fails to be fully informative since it generically refers to the DPO: there is no name, postal address, etc.

2. Clarity of expression: The policy should be framed in an understandable and precise language. For CLAUDETTE, they defined 4 main indicators of vagueness:

INDICATOR	LANGUAGE QUALIFIERS
1. Conditional Terms The performance of a stated action or activity is dependent on a variable trigger	<i>depending, as necessary, as appropriate, as needed, otherwise reasonably, sometimes, from time to time, etc.</i>
2. Generalization i.e. terms that vaguely abstract information practices using contexts that are unclear. Action(s)/Information Types are vaguely abstracted with unclear conditions.	<i>generally, mostly, widely, general, commonly, usually, normally, typically, largely, often, primarily, among other things, etc.</i>
3. Modality It includes modal verbs, adverbs and non-specific adjectives, which create uncertainty with respect to actual action; it includes whether an action is possible. Modality does not include whether an action and/or activity is permitted. Modality mainly refers to the ambiguous possibility of action or event.	<i>may, might, could, would, possible, possibly, etc.</i>
4. Non specific Numeric quantifiers Which create ambiguity as to the actual measure	<i>certain, numerous, some, most, many, various, including (but not limited to), variety, etc.</i>

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

Type of clause	Symbol
Processing of special categories of personal data (e.g. health, sex life, political opinions, religious beliefs, etc.)	<sens>
Consent by using	<cuse>
Take or leave it approach	<tol>
Third party data transfers	<tp>
Policy change	<pch>
Transfer of data to third countries	<cross>
Processing of children's data	<child>
Licensing data	<lic>
Advertising	<ad>
Any other type of consent	<C>

With CLAUDETTE, they are trying to design a service for Assessment of Privacy Policy, but this is way more complex than Analysis of Terms of Services and so they have not obtained super results in terms of performance so far.

Further steps for CLAUDETTE

- Experimenting new method for privacy policies
- Multilingualism (The Claudette german version)
- Empowerment through transparency:
 - Linguistic transparency,
 - Provide explanations opening black box AI Systems

WEB-CRAWLER

It is another tool developed for automatic privacy policy monitoring.

Two types of monitoring:

- Checking the date on the document
- Comparison of the content with the previously saved version

Earnings reports by email.

17/05/2021 - INTELLIGENT WEAPONS (Sartor)

The concept of Autonomy

Intelligent Weapons are also called *Autonomous Weapons*.

What is the meaning of “autonomy”? These are all possible definitions:

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

Problem of stating if an agent is autonomous:

- If the standard to state Autonomy is too high (namely all cognitive capacities of humans are required), no artificial entity is autonomous;
- If it is too low, all algorithms are autonomous.

NB: autonomous behaviour of a system in its complex may emerge from the interaction of lower level non-autonomous or autonomous elements.

In conclusion... from *“The autonomy of technological systems and responsibilities for their use”* by Sartor and Omicini, we consider Autonomy as a scalable capacity, merging three dimensions:

1. independence
2. cognitive skills
3. teleonomic cognitive architecture

1. Independence (and Independence within a socio-technical system)

A technological device, within a system, is independent to the extent that it is able to accomplish on its own, without external interventions a high level task. Examples:

- landmine
- airplanes’ collision-avoidance system or autopilot
- ...

Higher grade of autonomy could be the Independence within a socio-technical system: an integrated combination of human, technological and organizational components. Examples:

- airplane
- ...

2. Cognitive skills (and Cognitive Delegation)

An autonomous system engages in high-level cognition (involving the ability to discriminate facts, actions or outcomes) using its own abilities in one or more of the following ways:

- acquisition and classification of input data
- information analysis to extract further information from input data
- action selection, construction of plans of actions

- implementation of strategies

Of course this constitutes the main aspect to differ from a mere trigger-mechanism (like a landmine, which is only “independent”).

In order to state an agent has cognitive skills it must exploit automation for some cognitive task it needs to achieve:

- acquisition and classification of input data (input data, noise reduction, filtering, etc.)
- information analysis (compute expected flight trajectories or possible encounters, alert operator of possible risks, etc.)
- decision and action selection (suggestion, list of options, take action, etc.)
- plan implementation and monitoring (flying according to the established route, monitoring projective, etc.)

A further level of autonomy could be gathered by having Autonomous Cognitive Delegation:

- the delegator decides to delegate choices instrumental to the purpose achievement to the cognitive skills of the delegatee system (ex: flying aircraft, target-engager, etc.)
- the delegator does not know and thus does not intentionally pre-select what the delegated system will choose to do in future situations (ex: how to fly, what particular target to engage, etc.)

An example: BAE System Taranis¹², an autonomous drone-system, even if there is a human-in-the-loop presence, “pressing the button” to confirm the hit.

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

Humans may remain in the loop while technological devices execute the larger share of the cognitive functions involved in the performance of the task

3. Teleonomic¹³ cognitive architecture

Useful definitions:

Adaptiveness = a system that can change its patterns of behaviour to better achieve its purpose according to the environment in which it operates by changing its internal states as the environment changes.

Cognitive Behavioural Architecture = adaptiveness (as auto-teleonomy) + teleonomy (purposiveness) + intentionality

The next step in being autonomous is when systems are teleological and so have explicit cognitive states (goals, beliefs, plans, intentions...).

These cognitive states are

- differently implemented than corresponding human mental states,
- simpler,

but

- performing the same basic functions (indicating objectives, tracking the environment and directing future actions, storing executable commitments...).

¹² https://it.wikipedia.org/wiki/BAE_Systems_Taranis

¹³ <https://en.wikipedia.org/wiki/Teleonomy> Teleonomy is the quality of apparent purposefulness and of goal-directedness of structures and functions in living organisms brought about by natural processes

Autonomous Weapons

There is no actual agreement about it world-wide ⇒ many different approaches across different countries.

In 2012 USA issued the *Directive on Autonomy in Weapons Systems* that touches autonomous and semi-autonomous weapon systems.

The difference between them is about **target selection**:

- so no problem with, for example, a drone flying autonomously or with having someone governing it from afar ⇒ they are not even considered autonomous weapons by the USA Government, they are considered *semi-autonomous weapons*;
- but, when the weapon also does autonomous target selection (namely, once it has been activated by humans, can autonomously select and engage targets without further intervention by a human operator), this is considered to be an autonomous weapons;
- the USA declared they do not have such weapons.

So the basic difference between autonomous and semi-autonomous is that the latter are intended to only engage individual targets or specific target groups that have been selected by a human operator.

On this line, autonomous weapons, with their ability to autonomously select targets, should only be used to apply non-lethal and non-kinetic force, plus they may engage with non-human targets (ex: intercepting missiles), while semi-autonomous weapons may be deployed for any purpose, including the exercise of lethal force against humans.

NB: target selection includes all aspects of decision making (acquisition and classification of input data, information analysis, decision and action selection, implementation of chosen strategy) which can be automated partially or totally

A critique to the USA distinction:

- non-autonomous target selecting... the machine probably still has the to select the particular object to engage with, the human operator only delimit a domain for the target!
- al alternative autonomous target selection is when human operator gives a description of the target and the weapon engage the target comparing it with it
- the last one gets closer to the idea of teleological ability of cognitive architecture, which are indeed autonomous weapons

Responsibility

The UN's agreement seeks to outlaw war, but as long as we have them, we have to ask ourselves what is the role of AI in these scenarios.

Kind of responsibility (at war):

- Functional responsibility (what failure caused unwanted harm)
- Blameworthiness (if the failure that caused harm involves a fault of a moral agent)
- Legal liability for tort (if the moral agent should be legally prosecuted)

International Humanitarian Law

IHL is the law that regulates the conduct of war, "*Jus in bello*". It is a branch of international law which seeks to limit the effects of armed conflict by protecting persons who are not

participating in hostilities, and by restricting and regulating the means and methods of warfare available to combatants.

There are three fundamental Principles protecting civilians, which are basically important factors in assessing military force legitimation:

- NECESSITY
 - legitimate wars (by the UN), according to “*Jus Ad Bellum*”, when the country is brought at war for defensive necessity
 - there are some other contexts in which it is admitted to engage war, in order to fight for human rights that have been violated (*humanitarian wars*)
 - and then there are those military activities considered to be outlaw by the UN that are the supremative and aggressive ones
- DISTINCTION - “*belligerents must distinguish between combatants and civilians*”
 - and any harm to civilians has to be strictly related to the military goal
 - ex: nazi lagers were condemned; bombing by the Alleis over german cities and civilians were condemned even if the Alleis were reacting to previous attacks, etc.
- PROPORTIONALITY
 - the harm caused to civilians has to be proportionate to the military goal pursued in the war, avoiding excessive harm in relation to the concrete and direct military advantage anticipated

FACT: there are some prohibition according to the IHL that has been successful, like the banned chemical weapons after their effect in the Great War

AI at war

NB: it is very difficult to discriminate between technologies developed for war scope vs other scopes (example: civil airplane autopilot and war drone autopilot, facial recognition in civil purpose or in war context, etc.) therefore it is difficult to prohibit the development of war AI engines.

Big problem: impossibility of defining responsibility and of attributing moral responsibility and legal liabilities to anyone for certain harms!

Sartor’s conclusion: it is not realistic to think of a general ban of AI-based weapons, so what he would be concerned of is that weapons and military forces must respect IHL (so hybrid systems could probably be the “best” realistic case?).

24/05/2021 - ETHICS OF FILTERING (Loreggia)

Introduction

Filtering = any act of stopping, banning or removing any type of content

Moderation = active governance of platforms meant to ensure interactions among the users that are: productive or pro-social or Lawful

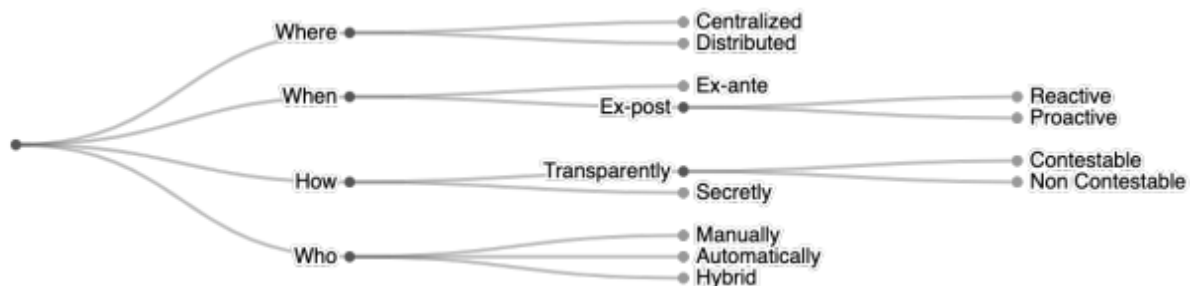
Digital Services Act (DSA) ⇒ regulation of digital services + online platforms

User-generated content ⇒ enable users to express themselves + create, transmit or access information and cultural creations + engage in social interactions.

Why filtering?

- To prevent unlawful and harmful online behaviour
- To mitigate the effect of unlawful and harmful online behaviour
- To facilitates cooperation
- To prevents abuse

Taxonomy



Where

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

When

- Ex-ante filtering: applied before the content is made available on the platform
- Ex-post filtering: applied to the content that is already accessible to the platform's users:
 - Reactive filtering: takes place after the issue with an item has been signaled by users or third parties
 - Proactive filtering: takes place upon initiative of the moderation system, which therefore has the task of identifying

How

- Transparent filtering: provides information on the exclusion of items from the platform
 - Contestable filtering: when 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: when there is no remedy available to the uploaders
- Secret filtering: not providing any information about the operation

Who

- Manual filtering: performed by humans
- Automated filtering: performed by algorithmic tools
- Hybrid filtering: performed by a combination of humans and automated tools (first go is automated then there is a second check made by humans)

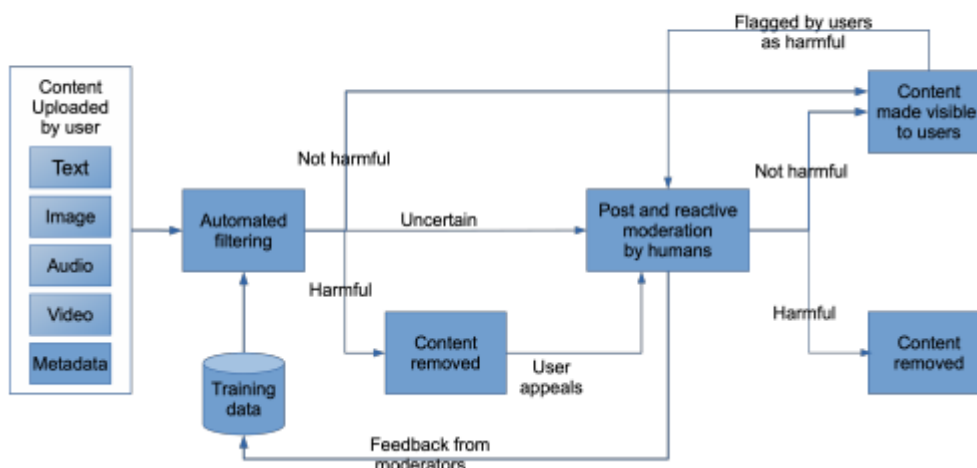
Different media

An aspect that makes everything more complicated is that the media whose content needs to be checked and filtered can be of various kinds. It could be audio, text, images and any combination of these, and of course depending on the kind of media the techniques to apply filtering change!!

Some example of techniques:

- over text, blacklisting, in order to find unwanted expressions
- materials subject to copyright policies are registered in a sort of database. When someone uploads something this content could be checked with the copyrighted database and if there appears to be unlawful material this can be filtered. The comparison between the uploading material and the database is done according to hashing or fingerprint checking and so on
- but the methodologies just mentioned may be fooled very easily by putting small differences in the content. This is why multiple AI techniques have been implemented to identify unwanted images, or combinations of text and images, and to translate spoken language into text, for example, in NLP domain, it is possible to address meaning and context.

Filtering process - How it works





Facebook banned Neptune statue photo for being 'explicitly sexual'

ML/AI techniques are statistical models that lack “common sense”, so they are prone to errors... which is why Post and reactive moderation by humans is needed.

It is even more complicated to understand differences and therefore proper behaviour in filtering in some cases...

For example:

- information and testimonials about a civil war (Syria) should be public,
- while images of terroristic, efferal actions should not be shared!



Regulations - Santa Clara Principles

Santa Clara University's High Tech Law Institute organized the “Content Moderation & Removal at Scale” conference and Eric Goldman supported the convening of the workshop that resulted in the document of [Santa Clara Principles](#).

These principles are meant to serve as a starting point, outlining minimum levels of transparency and accountability that we hope can serve as the basis for a more in-depth dialogue in the future.

The document's principles:

1. Companies should publish the **numbers** of posts removed 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.

Downstream studies show that **these principles got companies to be more transparent.**

Issues concerning Filtering

- Filter bubbles (avoiding some kind of content) and Echo chambers (repeating the same content over and over again), which end up building an environment of fragmented opinions where users get manipulated into approaching that environment
- Censorship
- Fake News

31/05/2021 - FRAMEWORK FOR ETHICAL PRINCIPLES “AI Ethics at IBM: From Principles to Practice” (Francesca Rossi, AAAI President)

AI limitations

- narrow AI (solves well specific problems but it gets way more complicated when a broader context is proposed)
- needs a lot of resources (data and computing power)
- lack of robustness and adaptability (ex: putting some noise and get misclassification)
- ethical issues



AI Ethics

This is a multidisciplinary field of study aimed at optimizing AI's beneficial impact while reducing risks and adverse outcomes.

So, how to design and build AI systems that are aware of the values and principles to be followed in the deployment scenarios?

To achieve this, it is necessary to identify, study, and propose technical and nontechnical solutions for ethics issues arising from the pervasive use of AI in life and society.

AI needs to be NEUTRAL.

Main AI Ethics issues

- AI needs data (many times personal data!)
 - related issue: Data privacy and governance
- AI is often a black box
 - Explainability and transparency
- in many cases, AI can make or recommend decisions to humans
 - Fairness (⇒ social justice and lawfulness) and value alignment
- AI is based on statistics and has always a small percentage of error (that we can never get rid of completely)
 - Who is accountable if mistakes happen?
- in many cases, AI can profile people and manipulate their preferences
 - Human and moral agency
- AI is very pervasive and dynamic
 - Larger negative impacts for tech misuse (by humans)
 - Fast transformation of jobs and society
- Bad use of the technology (ex: Autonomous weapons and mass surveillance)
 - vs Good use (ex: UN Sustainable Development Goals)

The issue of Fairness

- Individual vs group fairness:
 - similar individuals should receive similar treatments or outcomes, vs
 - groups defined by protected attributes should receive similar treatments or outcomes
- Context-dependent definition(s) of fairness
- Acceptable bias threshold
- When to detect bias:
 - training data or learned model

The issue of Explainability

This aspect is even more complicated... and is so necessary (even the GDPR states that a data subject has the right to obtain a meaningful explanation about the logic involved in deciding what to do with his/her personal data).

The issue of Profiling and Manipulation

- From actions to profiles
 - Like, text, images, follow, ...
- AI can infer our preferences, and use them to advertise products that we probably like
 - Easier if our preferences are bipolar

The issue of Impact on the workforce

Many jobs will disappear, and many others will be created. All jobs will change. Even if now it is difficult to imagine the new jobs of the future, just think that in the last century 90% of the population worked in agriculture while now only 2% does.

IBM and its ethical approach

IBM is 110 years old. Started out as a hardware and software company (they invented the personal computer) and nowadays they mostly create enterprise AI, so AI solutions for other companies (ex: Banks and financial institutions, Governments, Airports, Hospitals, etc.) and further research (ex: IBM Deep Blue (1997), IBM Watson (2011), Project Debater (2020), quantum computers, etc.)

IBM Principles of Trust and Transparency (2017)

- The purpose of AI is to augment human intelligence (and do not replace it!!) ⇒ this is why they deliver enterprise solutions to clients and deepen research
- Data and insights belong to their creator ⇒ they do not reuse data and solutions across clients
- New technology, including AI systems, must be transparent and explainable

...so what does it mean to TRUST a decision made by a machine?

- it has to be accurate and respect privacy
- it has to be fair (no discriminatory decisions)
- it has to be explainable and transparent (no black box)
- it has to be robust

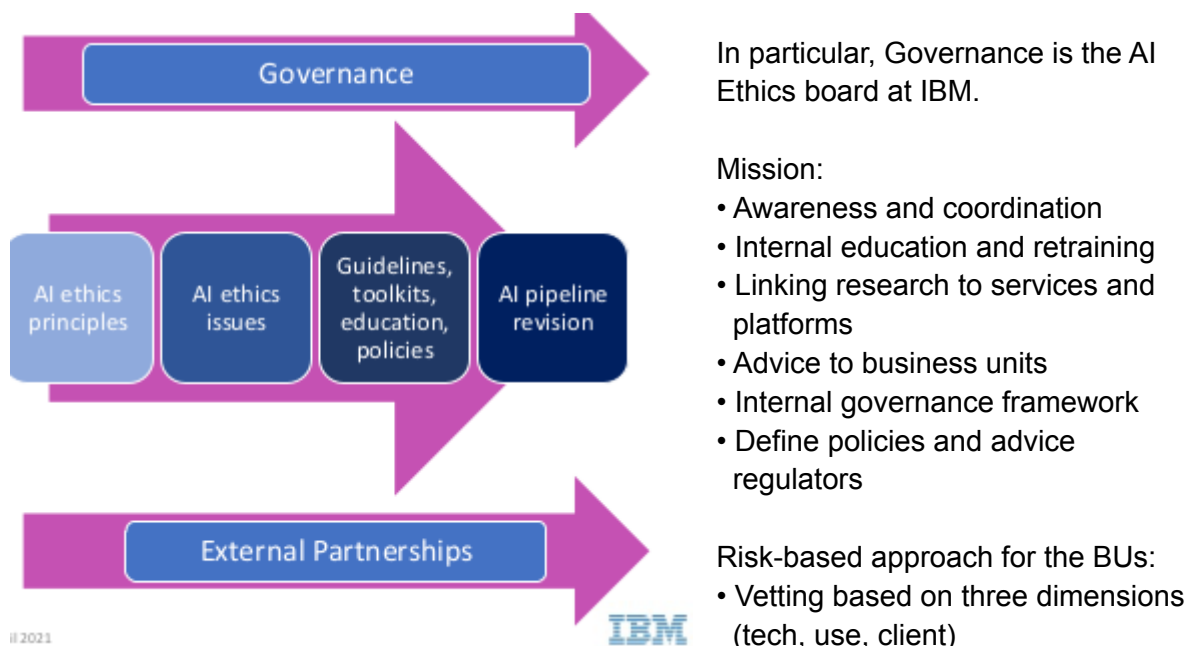
...and what does it mean FAIR at IBM?

- Technical solutions to detect and mitigate AI bias:
 - Research work
 - Watson OpenScale
 - Open-source libraries: AI fairness 360
- Developers' education and training
 - AI bias education modules for all IBMers
 - Developers' awareness material
 - Revised methodologies for the AI pipeline
 - Adoption strategies
 - Governance frameworks
 - Consultations with all stakeholders
 - Design thinking sessions

...and what does it mean TRANSPARENT at IBM?

- AI factsheet
 - Transparency by documentation
 - Design a development choices
 - Not just a checklist
 - Self-assessment and beyond
- Useful to
 - Developers
 - Clients
 - Users regulators/auditors
- Aligned with EC High Level Expert Group on AI self-assessment list (ALTAI)
- AI factsheet 360 (available to everybody to get familiar to all the previous concept)

From principles to practice: a multi-dimensional space



While External Partners could be Academia, Companies, Governments, Civil society, organizations... Multi-disciplinary and multi-stakeholder!!

IBM research is not just AI...

Other domains are:

- Neurotechnologies
 - Huge potential for healthcare
 - Reading/writing neurodata
 - Additional issues around privacy, agency, and identity
- Quantum computing
 - How to responsibly use such a huge computing power?

...and it is very easy to understand that all of these areas have a lot to do with ethical principles and human rights.