

# ETHICS GUIDELINES FOR TRUSTWORTHY AI

## By the High-Level Expert Group on Artificial Intelligence

Giovanni Sartor



# The document

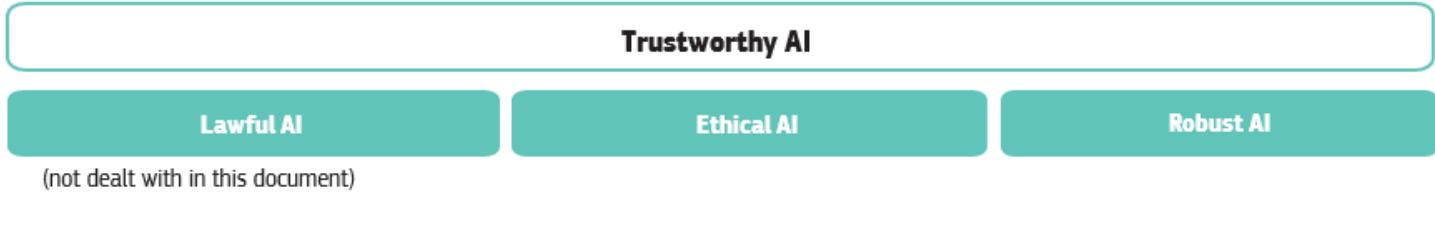
- Prepared by the High-Level Expert Group on Artificial Intelligence set up by the European Commission in June 2018.
- made public on 8 April 2019.
- available online (<https://ec.europa.eu/digital-single-market/en/high-level-expert-group-artificial-intelligence>).
- It is a good example of the many documents on ethics of AI published so far

# The idea of trustworthy AI

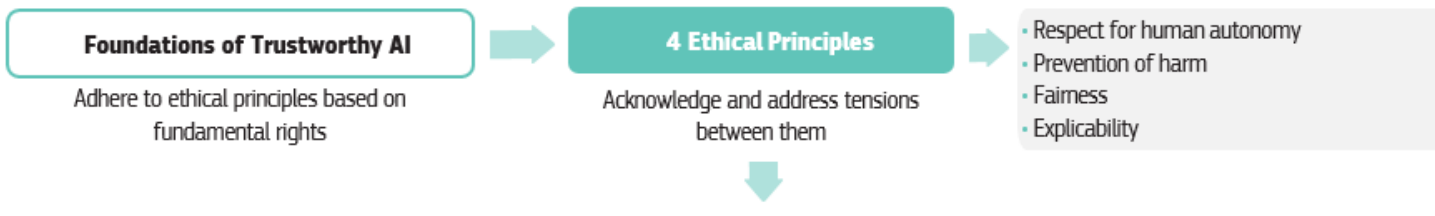
- AI should be
  - Lawful, complying with all applicable laws and regulations
  - Ethical, ensuring adherence to ethical principles and values
  - Robust, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm
- This requirements should be met throughout the system's entire life cycle
- Question. Can you think of examples of unlawful, unethical or non-robust uses of AI?

# Framework for Trustworthy AI

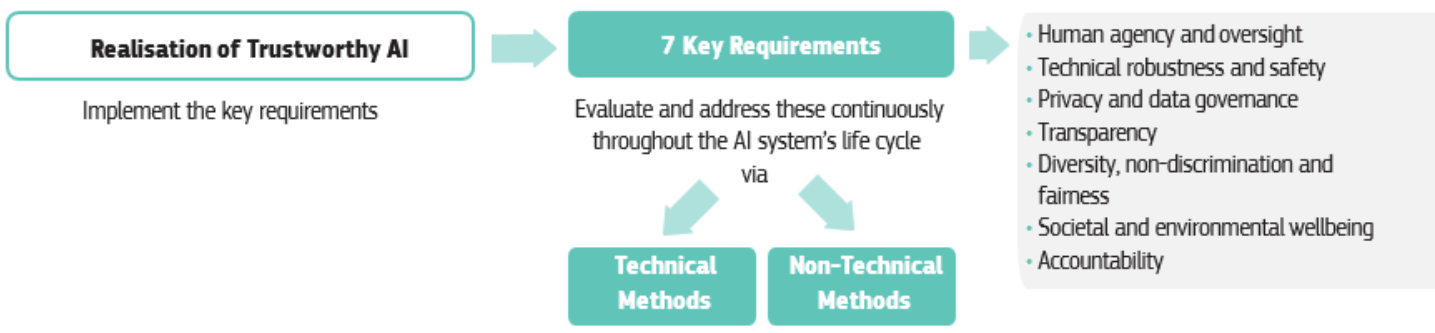
INTRODUCTION



CHAPTER I



CHAPTER II



CHAPTER III





# Chapter 1: Ethical principles

- Develop, deploy and use AI systems in a way that adheres to ethical principle :
  - respect for human autonomy,
  - prevention of harm,
  - fairness and
  - explicability.
- Acknowledge and address the potential tensions between these principles.
- Pay particular attention to
  - situations involving more vulnerable groups such as children, persons with disabilities and others that have historically been disadvantaged or are at risk of exclusion, and
  - situations which are characterised by asymmetries of power or information, such as between employers and workers, or between businesses and consumers.
- Acknowledge that, while bringing substantial benefits to individuals and society,
  - AI systems also pose certain risks and may have a negative impact including impacts which may be difficult to anticipate, identify or measure (e.g. on democracy, the rule of law and distributive justice, or on the human mind itself.)
  - Adopt adequate measures to mitigate these risks when appropriate, and proportionately to the magnitude of the risk.

# Chapter II: guidance of realisation trustworthy AI

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

# Chapter II: guidance of realisation trustworthy AI (continues)

- Foster research and innovation
  - to help assess AI systems and to further the achievement of the requirements; disseminate results and open questions to the wider public, and systematically train a new generation of experts in AI ethics.
- Communicate, in a clear and proactive manner, information to stakeholders about the AI system's capabilities and limitations,
  - enabling realistic expectation setting, and about the manner in which the requirements are implemented. Be transparent about the fact that they are dealing with an AI system.
- Facilitate the traceability and auditability of AI systems
  - , particularly in critical contexts or situations.
- Involve stakeholders throughout the AI system's life cycle.
  - Foster training and education so that all stakeholders are aware of and trained in Trustworthy AI.
- Be mindful that there might be fundamental tensions between different principles and requirements.
  - Continuously identify, evaluate, document and communicate these trade-offs and their solutions.

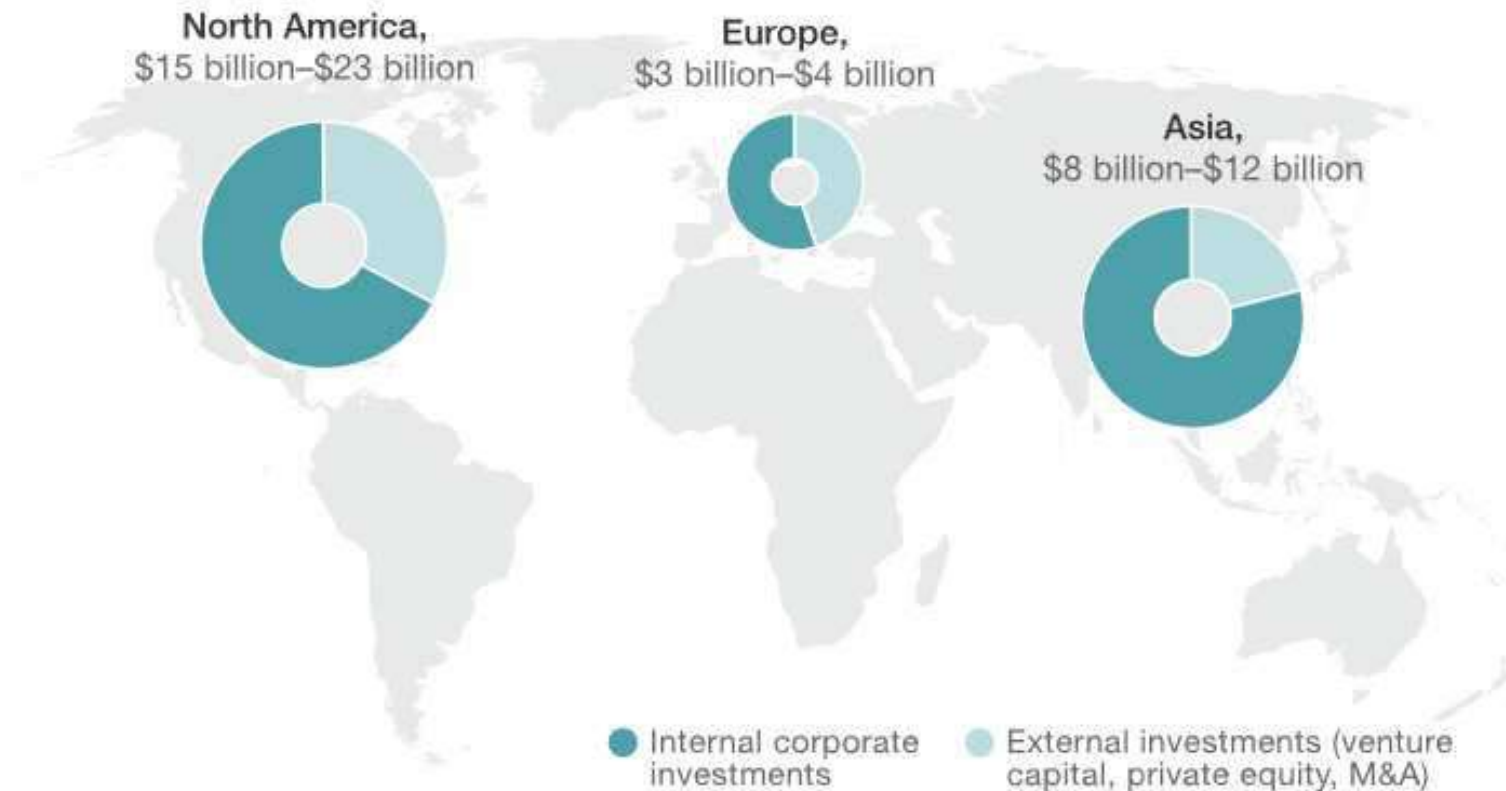
# Chapter III: Trustworthy AI assessment

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

# The Commission's approach to AI

- Communications 25 April 2018 and 7 December 2018 (COM(2018)237 and COM(2018)795). Three pillars:
  - (i) increasing public and private investments in AI to boost its uptake
  - (ii) preparing for socio-economic changes, and
  - (iii) ensuring an appropriate ethical and legal framework to strengthen European values.
- <https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-237-F1-EN-MAIN-PART-1.PDF>
- <https://ec.europa.eu/transparency/regdoc/rep/1/2018/EN/COM-2018-795-F1-EN-MAIN-PART-1.PDF>

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



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

# Human-centric AI

- commitment to the use of AI in the service of humanity and the common good, with the goal of improving human welfare and freedom.
- Maximise the benefits of AI systems while at the same time preventing and minimising their risks.

# Ethics vs law

- Ethics: norms indicating what should be done, with regard to all interests at stake
  - Positive ethics: norms shared in a society (possibly including ideas of social hierarchy, gender roles, etc.)
  - Critical ethics: norms that are viewed as most appropriate, or rational
- Law: norms that adopted through institutional processes and coercively enforced.



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

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

# AI should be lawful

- It should comply with
  - EU primary law (the Treaties of the European Union and its Charter of Fundamental Rights),
  - EU secondary law (regulations and directives, such as the General Data Protection Regulation, the Product Liability Directive, the Regulation on the Free Flow of Non-Personal Data, anti-discrimination Directives, consumer law and Safety and Health at Work Directives),
  - UN Human Rights treaties and the Council of Europe conventions (such as the European Convention on Human Rights),
  - Laws of EU Member State laws (Italian law).
- Laws can be horizontal or domain-specific rules (e.g., on medical devices)
- Issue: Can you think of a horizontal law covering all AI applications?

# Foundations of trustworthy AI

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



# Foundation: (Ethical) fundamental rights

- Respect for human dignity. Human dignity encompasses the idea that every human being possesses an “intrinsic worth”
- Freedom of the individual. Human beings should remain free to make life decisions for themselves: including (among other rights) protection of the freedom to conduct a business, the freedom of the arts and science, freedom of expression, the right to private life and privacy, and freedom of assembly and association.

# Foundation: (Ethical) fundamental rights

- Respect for democracy, justice and the rule of law. AI systems must not undermine democratic processes, human deliberation or democratic voting systems, due process and equality before the law
- Equality, non-discrimination and solidarity - including the rights of persons at risk of exclusion. In an AI context, equality entails that the system's operations cannot generate unfairly biased outputs. (GS: we need to understand what this means)
- Other citizens' rights the right to vote, the right to good administration or access to public documents, and the right to petition the administration

# Ethical principles (based on human rights)

- (i) Respect for human autonomy
- (ii) Prevention of harm
- (iii) Fairness
- (iv) Explicability

# Respect for human autonomy

- Humans interacting with AI systems must be able to keep full and effective self-determination over themselves, and be able to partake in the democratic process.
  - AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans.
  - they should be designed to augment, complement and empower human cognitive, social and cultural skills.
  - The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. T
  - This means securing human oversight over work processes in AI systems, supporting humans in the working environment, and aiming for the creation of meaningful work.

# The principle of prevention of harm

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



# The principle of fairness

- Substantive dimension
  - ensuring equal and just distribution of both benefits and costs, and
  - ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation.
  - Promoting equal opportunity in terms of access to education, goods, services and technology.
  - Never leading to people being deceived or unjustifiably impaired in their freedom of choice.
  - AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives
- Procedural dimension.
  - ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them
    - In order to do so, the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.

# The principle of explicability

- To ensure contestability
  - processes need to be transparent,
  - the capabilities and purpose of AI systems openly communicated, and
  - decisions – to the extent possible – explainable to those directly and indirectly affected.
- An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible.
  - other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required, provided that the system as a whole respects fundamental rights.
  - the degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate.<sup>3</sup>

# Tensions between the principles

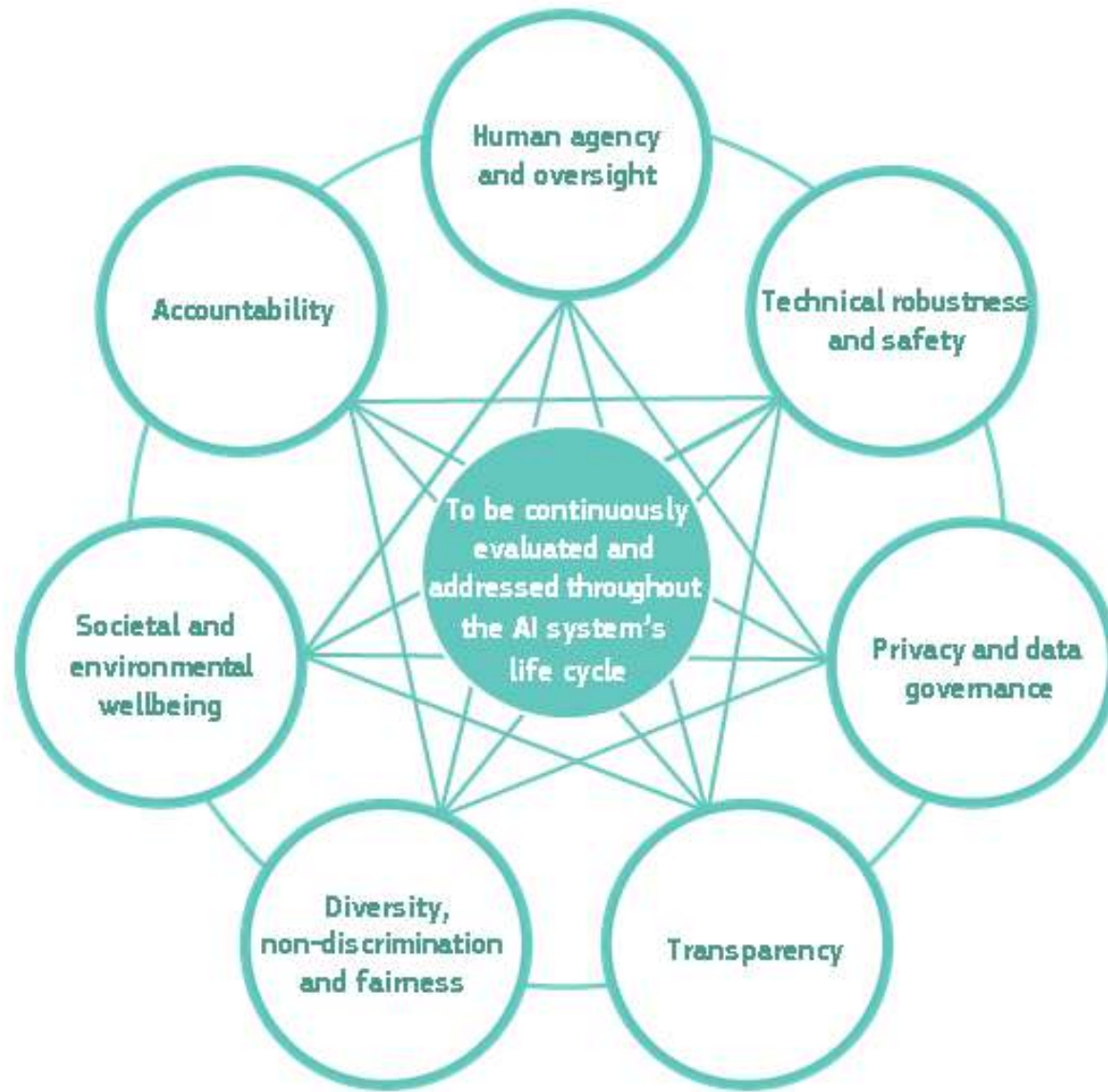
- Methods of accountable deliberation to deal with such tensions should be established.
  - Conflicts between prevention of harm and human autonomy
  - Also between welfare and security?

# Requirements of Trustworthy AI

- 1. Human agency and oversight
  - Including fundamental rights, human agency and human oversight
- 2 Technical robustness and safety
  - Including resilience to attack and security, fall back plan and general safety, accuracy, reliability and reproducibility
- 3 Privacy and data governance
  - Including respect for privacy, quality and integrity of data, and access to data
- 4 Transparency
  - Including traceability, explainability and communication

# Requirements of Trustworthy AI (continues)

- 5 Diversity, non-discrimination and fairness
  - Including the avoidance of unfair bias, accessibility and universal design, and stakeholder participation
- 6 Societal and environmental wellbeing
  - Including sustainability and environmental friendliness, social impact, society and democracy
- 7 Accountability
  - Including auditability, minimisation and reporting of negative impact, trade-offs and redress.



# Human agency and oversight

- AI systems should support human autonomy and decision-making. Therefore they should support
  - Fundamental rights
    - Human rights assessment
  - Human agency.
    - Users should be able to make informed autonomous decisions regarding AI systems.
  - Human oversight.
    - Human oversight helps ensuring that an AI system does not undermine human autonomy or causes other adverse effects (human-in-the-loop (HITL), human-on-the-loop (HOTL), or human-in-command (HIC) approach + public controls)
  - Technical robustness and safety
    - AI systems be developed with a preventative approach to risks and in a manner such that they reliably behave as intended while minimising unintentional and unexpected harm, and preventing unacceptable harm.

# Human agency and oversight (continues)

- Resilience to attack and security
  - AI systems, should be protected against vulnerabilities that can allow them to be exploited by adversaries
- Fallback plan and general safety
  - AI systems should have safeguards that enable a fallback plan in case of problems
- Accuracy
  - AI systems should have the ability to make correct judgements, for example to correctly classify information into the proper categories, or its ability to make correct predictions, recommendations, or decisions based on data or models.
- Reliability and Reproducibility .
  - The results of AI systems should be reproducible, as well as reliable.



# Privacy and data governance

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

# Transparency

- This requirement is closely linked with the principle of explicability
  - Traceability.
    - The data sets and the processes that yield the AI system's decision, should be documented
  - Explainability.
    - The technical processes of an AI system and the related human decisions should be explainable
  - Communication.
    - Humans have the right to be informed that they are interacting with an AI system.

# Diversity, non-discrimination and fairness

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

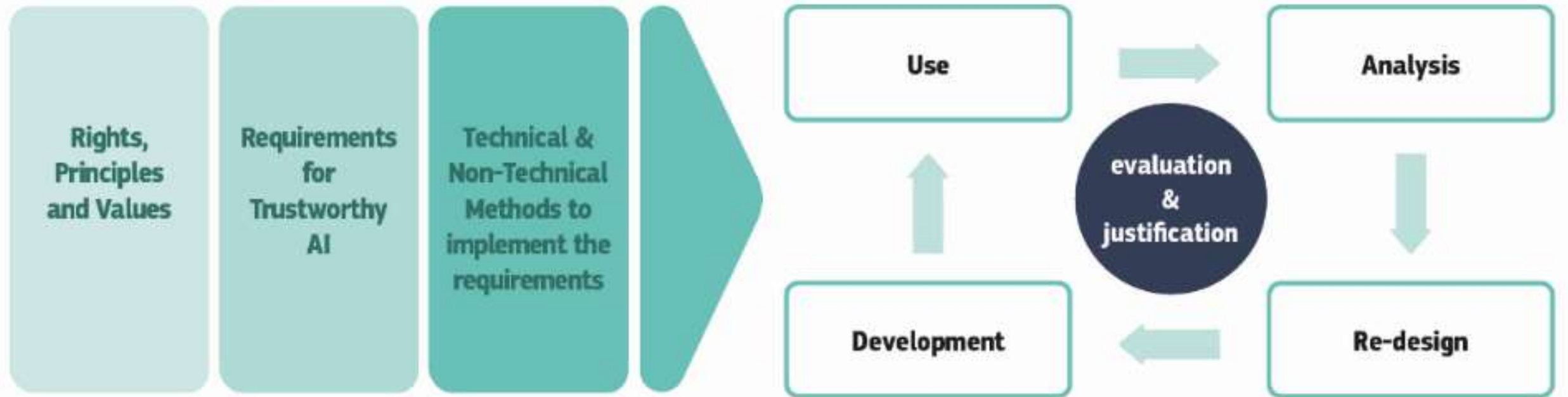
# Societal and environmental well-being

- The broader society, other sentient beings and the environment should be also considered as stakeholders throughout the AI system's life cycle.
  - Sustainable and environmentally friendly AI
    - Measures securing the environmental friendliness of AI systems' entire supply chain should be encouraged.
  - Social impact.
    - The effects of these systems on individuals, groups and society must therefore be carefully monitored and considered.
  - Society and Democracy.
    - Take into account AI's effect on institutions, democracy and society at large

# Accountability

- Ensure responsibility and accountability for AI systems and their outcomes
  - Auditability
    - Enablement of the assessment of algorithms, data and design processes
  - Minimisation and reporting of negative impacts
    - The ability to report on actions or decisions that contribute to a certain system outcome, and to respond to the consequences of such an outcome, must be ensured.
  - Trade-offs
    - Trade-offs should be addressed in a rational and methodological manner within the state of the art
  - Redress.
    - Accessible mechanisms should be foreseen that ensure adequate redress

# Technical and non-technical methods to realise Trustworthy AI



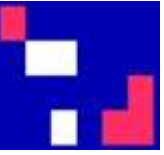
# Questions and suggestions

- Questions

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

- Suggestion

- Read all the document!
- Read also

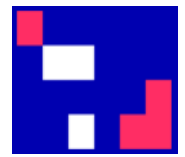


# Thanks for your attention

- [Giovanni.sartor@Unibo.it](mailto:Giovanni.sartor@Unibo.it)







# For a Science-oriented AI & not servant of the business

---

**Cristiano Castelfranchi**

*Institute for Cognitive Sciences and Technologies – Roma*



# We (will) live in an **AUGMENTED** and **MIXED** world/“**reality**”

Not just “Onlife” on the WEB (Floridi); living “connected”,  
but *in a new material world/reality*.

We will **act** in the Virtual for changing the Real;  
and vice versa.

We are “**present**” where we “are not”;  
we see and **act** where we “are not”.

And “somebody”, which is not “here”,  
will in fact *act* here and be “**present**” here.

We (will) live in a **HYBRID Society**,

a mix of human intelligences and 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**.

You are **social engineers**. Are you aware of?

You are **social engineers**. Are you aware of?

---

I will focus on

**A)** The importance of the **SCIENCE** side of AI;

**B)** 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.

A

**For a Science-oriented AI**

**The pleasure of research** (also in AI) should primarily be knowledge, discover, ideas, not just application and technology.

AI has a too strong “technological identity” more than a SCIENCE identity.

AI provides conceptual and cognitive (formal) instruments for modeling and thus **UNDERSTANDING** minds, intelligences, action and interaction, emotions, organization, ...knowledge.

AI should be proud of the crucial contribution it gave to the **scientific revolution** in XX and XXI centuries due to **the impact of the *Science of the Artificial* on behavioral and social science** (Herbert Simon)

## **In science**

the economic, social, technical outcomes should mainly be “collateral/unintentional” effects.

There must obviously be a research not generically K-oriented (“basic”) but ***oriented to solve problems***, but also in this “applied” research *the priority is K, understanding, explaining, modeling..*

AI sometimes looks a bit perverted at the full service of business, for providing new market products:

the new richness, the new industrial capital (Google, Amazon, etc etc etc)



The *scientific* advantages of  
the ARTIFICIAL, **SYNTETIC APPROACH**  
to Mind and Society  
is **UNDERSTANDING by BUILDING and SIMULATING**

ISTC-CNR group exploited that in several domains. On language, on autonomy, cooperation, sociality, trust, emotions, norms, power, etc.

=====

**AI scientific models:**

- 1) for **modeling/explaining human & natural Intelligences;**  
(Grosz on Conversation & shared plans, Ferrari's cit. Winograd)
- 2) for **emulating** them;
- 3) for **creating new** intelligence  
and its theory ("General Intelligence")

Philosophers frequently claim that what AI and cognitive scientists are doing is to **“anthropomorphize” machines** (that cannot in principle really have “mind”, “intelligence”, “intentions”, etc. but just “simulate” them)

On this debate see for example Floridi and Sanders

It is exactly the other way around:  
what we are doing is  
to **“de-anthropomorphize”** such concepts,  
making them *no longer “anthropocentric”*  
*but more general and abstract*, and more  
clear, formalized, and “operationalized”. No  
longer common-sense “words”.

AI mission isn't just to acritically *buy concepts and theories from human and social sciences* and philosophy for **“applying”** them.

AI gives back a crucial contribution, not just “technological”, by **changing those concepts, models, and theories.**

Not only our environment and society will be hybrid and augmented, **but our brain and mind will be *augmented*, new cognitive power and new functions.**

**Our cognitive capabilities will not just be improved,  
but *changed*.**

It is not only matter of “mnemonic functioning”, externalized memory, Data access and processing, of “reading”, of “learning by doing”.

There will be a serious **evolution of our “*social cognition*”** in the Hybrid society.

In particular the WEB (“Minds on Line”) and Virtual reality will empower:

- >> “collective intelligence and problem-solving”,
  - >> “collective sense-making”,
  - >> “knowledge capital and sharing”,
  - >> “creativity”, and
- > a new “embodiment of our cognitive representations”
- > our perception of *space, time, intelligence*,... will be changed
- > an extremely “externalized/distributed cognition and mind”

One of the main functions of the brain is **integrating** and **augmenting** the perceived reality:

- > With the affordances, ..
- > With the past, the future (expectation, predictions, objectives,.....)

**B**

**The AI revolution:  
empowering whom?**



**ARTIFICIAL SOCIALITY?**

We – AI & MAS community - are responsible for the introduction of “Agents” as

★ “**autonomous**” (proactive, with initiative, with their own learning, reasoning, evolution, .. ) and “**social**”,

cooperating with human by following true

★ “**norms**” (but also – in case – violating them),

and critically adopting our goals (**not just**

★ “**executing**”), with *over-help, critical-help, .....*

And **this was a correct and unavoidable solution**, for a real “Intelligence” interacting with us and usable from humans.

We (ISTC group) are **not repented AT ALL**, of contributing to model

**ARTIFICIAL sociality**

on the contrary.... however

This obliges scientists to **become aware** of **possible appropriation** of their creations, of possible **unacceptable uses** of these instruments.

**Are we missing the control?**

**Not** of our Autonomous Agents, Robots, etc.

but of **their possible uses?**

Are we ready for the **ANTHROPOLOGICAL REVOLUTION** grounded on Intell Technologies and artificial mixed society?

Which also is

**an economic, social, and political revolution.**

Are there **DANGERS** in *living with* Artificially Intelligent Agents and Robots?

Being **replaced** (practically or *cognitively*) or supported and **guided** by them?

---

Are there **DANGERS** in AUGMENTING our INTELLIGENCE and changing COGNITIVE PROCESSING?

For the mass media, the main **PROBLEMS** are :

- *Privacy*
- *Security (on WEB, ... on access ..)*
- *Fake news, misinformation*
- *Hackers' attacks*
- *Anthropomorphism*
- *War and Artificial soldiers/arms*
- *Ethics **inside** Artificial creatures and algorithms*

For the mass media, the main **PROBLEMS** are :

- **War and Artificial soldiers/arms**

Subra Suresh, Carnegie Mellon's president, said **injecting ethical discussions into A.I. was necessary** as the technology advanced. While the idea of "Terminator" robots still seems far-fetched, **the United States military is studying autonomous weapons that could make killing decisions on their own...** —

Finally solved the problem of the poor general:



## Bertolt Brecht (1898-1956)

“General, your tank is a powerful vehicle  
It smashes down forests and crushes a hundred men.

*But it has one defect:*

*It needs a driver.*

General, your bomber is powerful.  
It flies faster than a storm and carries more than an elephant.

*But it has one defect:*

*It needs a mechanic.*

General, man is very useful.  
He can fly and **he can kill.**

**But he has one defect:**

**He can think.”**

---

Finally generals no longer need a (human) driver or mechanic!!

**The AI driver can think, yes; but we/generals can *decide and control*  
*HOW it will think! (?)***

# “Engineering **Moral Agents**” :

Dagstuhl Seminar etc.

“Imbuing robots and autonomous systems with **ethical norms and values** is an increasingly **urgent challenge**, given rapid developments in, for example, driverless cars, unmanned air vehicles (drones), and care assistant robots.”

- *implementation of moral reasoning and conduct in autonomous systems*
- **NOT just surveillance but INTERNALIZED values and control**



# The mass media' **PROBLEMS** are mainly:

- *Privacy*
- *Security (on WEB, ... on access ..)*
- *Fake news, misinformation*
- *Hackers' aJacks*
- *Anthropomorphism*
- *War and Artificial soldiers/arms*
- *Ethics **inside** Artificial creatures and algorithms*

## For me not less serious problems....

Putting aside the future of **WORK** in 4.0 economy!

**B1**

Is our Intelligent Technology research  
**ONLY BUSINESS ORIENTED**  
**just because it needs money?**

## Meeting of **the minds** for machine intelligence

**Industry leaders, computer scientists and students, and venture capitalists** gather to discuss *how smarter computers are remaking our world.*

Once a machine is educated, it can help experts make better decisions

... **savvy machines can help us evaluate (social) policies.** Etc...

Are **ONLY THESE THE RIGHT SUBJECTS/MINDS TO INVOLVE ?**

for discussing about ethical and political and social consequences of machine intelligence and hybrid society?

**What about other subjects to be involved** like: moral and political philosophers, social scientists, trade unions, social movements (like women movement, like “occupy Wall Street”,..), politicians, poor countries, etc.?

# MIT News

ON CAMPUS AND AROUND THE WORLD

Meeting of *the minds* for machine intelligence

Industry leaders, computer scientists and students, and venture capitalists gather to discuss *how smarter computers are remaking our world.*

Why alliance only between **academy,**  
**scientists,** and **capitalists** and **business men?**

Is this so OBVIOUS and UNDISPUTABLE in  
current culture to become **INVISIBLE?**

## Meeting of **the minds** for machine intelligence

**Industry leaders, computer scientists and students, and venture capitalists** gather to discuss *how smarter computers are remaking our world.*

Once a machine is educated, it can help experts make **better decisions**  
... **savvy machines can help us evaluate (social) policies.** Etc...

### **“Better” for whom?**

It is **not a “technical” problem**, but a political problem. “Better” for poor and powerless people/countries

**or** for dominating classes, lobbies, powers, countries?

**“Better” for whom?**

It is **not** a **“technical” problem**, but a political problem.

**WE want a “beneficial” AI, but... for whom??**

Do not assume that *if something is beneficial it is beneficial for everybody*.

In society there are serious contrasts of interest and goals. Thus if something is beneficial for X (that is favors his/her goals or interests) is noxious for Y.

If AI is subordinated to and beneficial for profit and business interests is NOT necessarily beneficial for workers.

If is Beneficial for dominant countries not necessarily is beneficial for poor and colonized countries.

>> For being **BENEFICIAL** AI should first choose on which side to be.

# AI can be **VERY** beneficial

- for DEMOCRACY,
- for good market, with reduced deception and manipulation;
- for social planning and decision, and political imagination, projects;
- or transparency and control, participation

**“AI for FREEDOM”** (JICAI-ECAI ‘18) **of PEOPLE!**

# New Research Center to Explore **Ethics of Artificial Intelligence**

By JOHN MARKOFF

**NYTimes** - NOV. 1, 2016

Carnegie Mellon University plans to announce on Wednesday that *it will create a research center that focuses on the ethics of artificial intelligence.*

The ethics center, called the **K&L Gates Endowment for Ethics and Computational Technologies**, is being established at a time of growing international concern about the impact of A.I. technologies.

That has already led to an array of **academic, governmental and private** efforts to explore a technology that until recently was largely the stuff of science fiction. ... Peter J. Kalis, chairman of the law firm, said the potential impact of A.I. technology on the economy and culture made it essential that **as a society** we make thoughtful, ethical choices about how the software and machines are used.



# “AS A SOCIETY”?

“an array of **academic,**  
**governmental** and **private** efforts”

AGAIN:

Why an alliance **only** between academy, scientists, and capitalists  
and business men, (and war powers)?

Is this so OBVIOUS and UNDISPUTABLE in current culture?

Is our Intelligent Technology research **ONLY BUSINESS**  
**ORIENTED** just because it needs money?

**B2**

**Hidden Interests  
&  
AWARENESS technology**

Security, Privacy, War, Ethics, .. are for sure very relevant issues, we have to reflect on,

**BUT not the most or the only relevant ones** from the moral and political point of view.

>> **Hidden interests, manipulation** of us (users and programmers), exploitation, ... **emptying democracy**, etc. are NOT less important.

Scientists have to **be conscious**

not just manipulated, unaware although genial servants of those forces and interests.

>> **Democracy is not a formal and misinformed voting ritual.**

**WE** have to foster a **real “intelligence”** (understanding) and **EMPOWERMENT** of people in/on the hybrid societies evolution.

Not only **improved and collective INTELLIGENCE**  
but

improved and **collective AWARENESS,**

which is a crucial form of “intelligence”,

Understanding **what we are doing** and **WHY**  
we are doing that; who is “nudging” us.

HELP in **RATIONAL DECISION MAKING**, (by revealing and correcting our rational & affective **BIASES**) is OK, but...

the real problem is not that “our” decision be fully efficient and **rational** (not misinformed or biased), but:

**in favor of whom?**

With **AWARENESS** of “**interests**” we are serving

Intelligent Agents and algorithms have to **help us to understand** *not only our Goals* and how to RATIONALLY decide (not misinformed or biased), but also to **understand**

**in favor of whom?**

**Our “finalities”/“aims” ,**

**which go much beyond our mental Goals.**

## Also the **Goals of our Agents and Robots**

Are they explicit, **transparent at least for us?**

---

(Ro)Bots & Agents should be *comprehensible* and *trustworthy*: they must be able to **EXPLAIN us**:

- **WHY** they do/did what they do/did;
- The **REASONS** and **MOTIVES** of their actions, decisions, or suggestions.

NOT showing us their “algorithm”!

This requires a **COGNITIVE MODEL** of “reasons” and “motives” for believing, and for goal processing and decision. (AI for **SCIENCE**)

Moreover: **the Goals of our Agents and Robots**

*serve to* ***FUNCTIONS***: external, not chosen and represented GOALS.

---

**Do they favor some interest?**

*Is this transparent for us?*

---

To which **VALUES** do they respond ?

Perhaps do not shared by us but at least clear! **Or obscure?**



# "INTERESTS" Theory

What is better for me and my goals but...  
... I do not understand or intentionally  
pursue them.

## **Tutelary Role** Theory:

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

In a lot of circumstances Agents will:

- **decide *for us*** (delegated or not by us),

**or**

- **give us recommendations or just a little push** (the celebrated liberal “**nudges**”) *like in marketing,*

**But.. in a TUTELARY ROLE ?**

Moreover:

**Who** is judging what is **better for me**, or for us?

Is this really “in *our* interest” or primarily in the  
INTEREST of financial and informational  
dominant powers?

Or (in many countries) of the political regime?

This holds also for more explicit influencing devices like

>> ***RECCOMENDER SYSTEMS***

which will **know us better than us.**

Will they give us recommendations and suggestions “in our INTEREST”, in a TUTELARY attitude, or will they follow market criteria

**just a *more effective, personalized advertising*?**

**On the side of the “user”?!**

**Or of the “seller” (of our data or of some good)?**

They will **decide "for us"**,

but... **AMBIGUOUS: "instead of" us**  
or also **"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?**

Moreover: “TUTELARY” doesn’t means  
“protecting me”

only caring of our “**individual**” “**personal**”  
**interests**, but also helping us to understand and  
take care of:

- of **Common interests** and possible collective subjects and communities and pressures;
- of hidden **conflicts of interests**;
- of the “**commons**”, of public goods and their relevance and respect (environment, energy, water, **public health**, ...)

**“AUGMENTED INTELLIGENCE”**

**also means**

**AUGMENTED SOCIAL AWARENESS:**

**HOW** does it work the **“INVISIBLE HAND”** (the god of liberalism)

**which organizes the emergent and  
“spontaneous” social “order”.**

# **PRESENCES, Agents, robots, ...**

It is **again** a matter of:

**Which political and moral values will they care of?**

not our “car driver”

but the “society drivers” !!

and **our life-navigator.**



**3**

# **The “Mouth of Truth” Algorithm**



Clearly we are developing **algorithms for ascertaining the “truth”** in that mess of data, of assertions, hoaxes, and news, which will be diffused and accessible through the WEB.

An Algorithm for deciding about *reliable sources*, *credible information*, what is “true” among so many different claims and data.

**There is no alternative on that. However:**

- On which base such algorithm will “ascertain what is true”? Only on the basis of reliable and convergent sources? Of their number and net? On direct or indirect access to the “fact”?
- Also on the basis of the “values” and on the sharing and acceptability of the values of the source?

Even for ‘official’ science: is it always capturing or saying the truth?

- And there will be **dogmatic truths** and **undisputable authorities**, like in any culture?



*The people will believe what  
the media tells them they believe.*

- George Orwell

- And **which culture and values** will be assumed as the “right” ones?

How will we allowed to distinguish between a conflict of values or of interests from a mere conflict between more or less credible data, more or less grounded, direct, controlled, reliable, ..?



3

**‘PRESENCES’**  
in our  
**MIXED REALITY and SOCIETY**

The autonomous and proactive intelligent entities will **become**  
**'presences' and 'roles'** in our **hybrid society** (human and  
artificial agent)  
and **mixed and augmented reality** (combined **virtual and**  
**'real', 'natural'** and **automatic/prosthetic** world).

Now the problem will be:

**are we able to manage these *autonomous*  
and *too informed and intelligent* agents?**

It is a matter of:

➤ **A)** Which **roles** will those material or immaterial, visible and invisible “entities” play in our life and environment?

(work with Ricci & Tummolini)



A) Which **roles**

Will they be

**our Guardian angel**  
with a **'tutelary' role?**

By helping, protecting and empowering us



A) Which **roles**

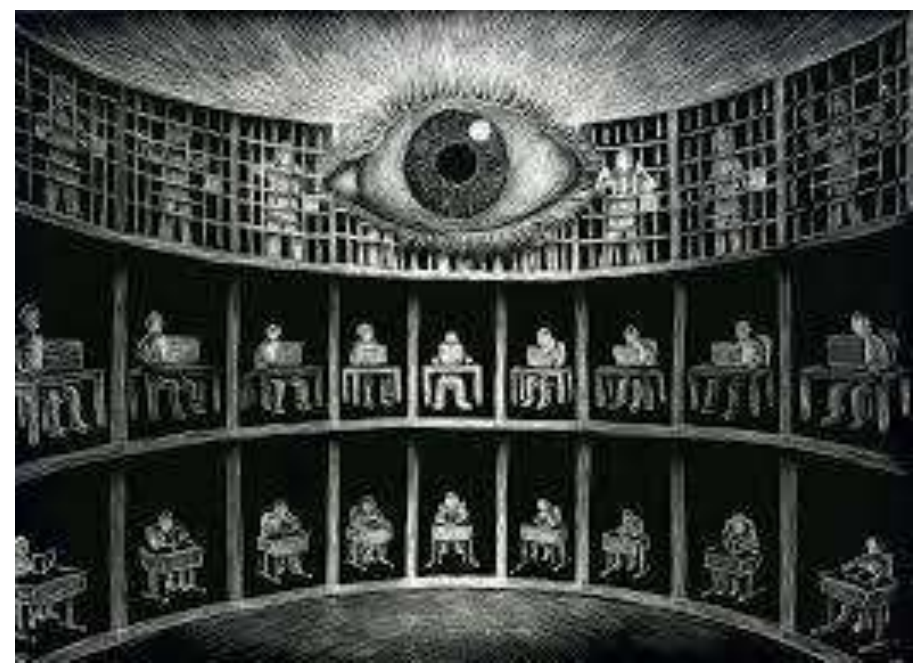


Or – less religiously – our Jiminy Cricket  
**(The Talking Cricket)**  
with its recommendations?

For example, we will not be the addressee and manager of our own  
“quantified self” and “lifeLog”.

# A) Which **roles**

Or our **supervisor** in the  
**ICT-Panopticon**  
we live in?



## The Prisoner and the Free



can't see the watcher

knows he's being watched

rounded back

all alone  
within a crowd of people



A) Which **roles**

or **our tempting Spirit**



A) Which **roles**

or our **tempting Devil** :

for the benefit of some  
marketing policy or monopoly,  
or the influencing and  
manipulating manager  
for hidden political  
or economic powers ?



# MIXED REALITY, MIXED BODY & MIND

Will we "incorporate", feel them as **parts of "us"**,  
our **"mental prosthesis"**?

Will we listen to that moral or rational "voice"  
as **our own** mental or consciousness voice  
**our (expanded) *SuperEgo***.

# MIXED REALITY, MIXED BODY & MIND

Or will our **Super Ego** be “externalized”?  
Not “me”.

Will we listen to "her" as to the voice of our mother,  
our teacher?

Or will we become “voice hearers”??

# MIXED REALITY, MIXED BODY & MIND

Both solutions will be probably there:

> The “social” one: **Externalized voices and Agents**

Our best friend; our sexual partners,..

&

> The “reflexively social” one: **an augmented internalized  
Self and Consciousness**



# PRESENCES

It is **again** a matter of:

➤ **B) Which political and moral values will they care of?**

not our “car driver” but the “society drivers” !!  
and **our life-navigator.**

They will **decide "for us"**,

but... **AMBIGUOUS: "instead of" us**  
or also **"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?**

**B3**

# **DIS**AGREEMENT TECHNOLOGIES

**A)** There is a **too strong ideology and rhetoric**  
**about society as cooperation, collaboration,**  
**common intent, collective advantages,.....**

how to reach convenient agreements and  
equilibrium, etc.

Moreover, **the web is (non accidentally) favoring a deviating political feeling: “we” against “them”** (governors, political caste, centralized powers).

This perception of “we” is completely misleading: there is no a “we” with common values and goals and interests, which has to be unified against the political power as such (in case against the real power (financial power) that has usurped the political power).

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

this is **the real conflict**

**(not “we” and “them”),**

and political activity and forces were supposed precisely to represent and protect those different social interests, and not just the “common” interest.

**Some conflict of interest or of value can be solved and reconciled in **a common interest**,**

but a large part of political/government decision is **not for a common advantage** (except reducing civil war), for a fair distribution;

it is for the prevalence or advancement of **the interests of a given group** (class, lobby, gender, view, ...) by reducing the powers of the others.

# The Need for Conflicts

## Conflicts: the presupposition of Democracy

Conflicts are **not just conflicts of views or opinions**, or due to different conceptions, information, reasoning.

**There are conflicts of "objective interests"**

the problem is conflicts between interests of group or classes, or conflicts between "private" interests vs. common interests, the "commons" and public goods.

Social conflicts in fact **do not have** a "verbal/cognitive" or a "technical" solution, just based on data and technical principles;

**they have a "political" solution;**

it is a matter of "power" and of prevailing interests and compromises (equilibrium, partitions/shares).



# The Need for Conflicts

## Conflicts: the presupposition of Democracy

### *No conflicts no democracy*

Democracy is not only a "response" to Cs and for moderating them; it would be a way of encouraging, growing (and solving).

Conflicts are not only to be governed, reduced, reconciled: they should even be *promoted* and this is in fact the role/function of specific forces and organizations, like trade-unions, parties, group of interests, associations, movements, etc. Crucial stakeholders of democracy, but also definitely responsible of the typical social, cultural, economic "progress" of western countries in the last centuries and now of the rest of the world.

Of course conflicts might be dangerous conducting us to fighting, violence, war, .. So it is true that societies and groups need "rules" for governing them, to avoid degeneration. Centralized state was one of these solutions: the state monopolizes violence; private or group violence is forbidden.

# Viva Conflicts!

Conflicts

with their **disagreements and agreements**

are thus the motor and principle of Democracy and of its possible effectiveness in changing society in favor of the submitted subjects, disadvantaged classes and groups, etc.

**Viva conflicts!**

# Democracy

Mark Twain is brilliantly right

*"If voting made any difference they wouldn't let us do it."*

But... the problem is much harder; it is not just a complot, is that we vote in a self-defeating way, and, in general, our collective stupidity.

**Might political "education" and education to "commons"  
& Digital society and participatory democracy  
be enough, and solve this "cognitive" and social problem?**

They will help. But

given the immediate local perception of the conflicting interests and competition and the *blindness to common interests* among different countries and poor classes and ethnic groups, and affiliation and identity feelings, conformism, and in-group vs. out-group psych, ... I have some doubt.

In a couple of centuries they will see.

- To *give voice to people* never in condition to protest, and to be listen to....
- *Making **conflicts** to **emerge** and become aware of, making express disagreement, making transparent which interests are hidden and prevailing, ...*

should be (in democracy) **one**  
**of the main tasks of intelligent social**  
**technologies.**

## B) “critical thinking”

Using WEB technologies for organizing “movements” it is OK; but **not so good**

**without promoting critical consciousness**

Not only by:

- Counteracting our *Confirmation Bias*;
- Counteracting our *tendency to gregariousness* and the “*bubble effect*” on the WEB

but by helping us to

**understand hidden powers, and also  
our prejudices.**

We need environments and Agents for learning and developing a **“critical thinking”** attitude; to manage our **cognitive and motivational biases**; etc.

To support us in argumentation and discussion, and in understanding the tricky arguments of the others.

To resist to the prevalence of “audience” against “quality”, of self-marketing and indexes against originality and quality; etc.

.... about propaganda, Academy, gender models, fanaticism, superstition, urban legends, ...

**We have impressive possibilities** with new intelligent and interacting technology, big data, etc.

They shouldn't be just used for selling and for dominating.

# Demystifying the Ideology of the NET

*NET interaction is perceived as non hierarchical, without superstructure and mediation, individually managed, spontaneous, thus “free”. Really and directly “democratic”.*

A neoliberal view and a wrong perception.

- There are **new Powers** beyond the WEB and its activity and information;
- Impressive **oligopolistic economic interests**
- **Influence, manipulation,**
- **Exploitation** of data, Exploitation of work

## DISAGREEMENT TECHNOLOGIES

### C) *anti-manipulation*

ICT and cognitive technologies are used for recognize our profile and interests but

**NOT for EMPOWERING US,**

**but**

in order to propose/*induce us to “buy” something* (goods, ideas, ..)

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

---

We need *anti-manipulation* AI technologies:

I would like to have not so much a **personal virtual or robotic psychotherapist or physiotherapist;**



# DISAGREEMENT TECHNOLOGIES

## C) *anti-manipulation*

I would like much more a **“life navigator”** in my main “social role” (ex. consumer!), but not a navigator saying “turn right, turn left”, “buy that; do not buy this” ...

But a **tutor, a trainer**, inducing me to understand and to reflect about why I’m oriented in that direction, I’m choosing that product; 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 great **REVOLUTION** of ICT, of digital monitoring and predicting (by simulation) and **BIG DATA** can give to society (to demos)

**a glass were to observe themselves and follow what it is happening.**

**A glass reflecting also what is invisible: hidden presences, the future (predictions for planning):**

**A GLASS OF the INVISIBLE**

The great **REVOLUTION** of ICT, of digital monitoring and predicting (by simulation) and **BIG DATA** can give to society (to demos)

**a glass were to observe themselves in the future:**

“The best way to build the future  
is to predict /simulate/imagine it”

(Nala Yak)

# The GLASS of the INVISIBLE

Not only "PRESENCES"; what is "not present" here, but can be virtually present for interaction, can act in this word and vice versa, etc.

But also a glass able to show what cannot be seen/understood: the **future, predictions**, the "emergent" order, and

**hidden phenomena and interests**

(for example, can I **see** who is now getting my personal data? And for what?

For whom am I **working for free**?)

# Can We Overcome Human **Alienation**?

Could we, by exploiting

> **collective, distributed, hybrid INTELLIGENCE**

and

> **BIG DATA**

and

> run-time feedbacks and information from **local stakeholders and intelligent sensors**

and

> **Computational LEARNING and PREDICTING**

and

> Computer (Agent-based) **Social SIMULATION**, and **VIRTUAL REALITY and SERIOUS GAMES**, etc.

could we

**MAKE VISIBLE the INVISIBLE  
HAND?**

and (partially) **GOVERN IT?**

# Can We Overcome our **Alienation**?

Will the **Leviathan** become

a giant *connected and informed community of agents,*

managing their Collective Power?

1. I'm skeptical about that (also for cognitive reasons)
2. I worry about possible **net-Demagogy**

***To See What Is (currently) Invisible:***

***Artificially Augmented Awareness***

**the real revolution of AI**

**(Including itself! It uses)**



In the Digital Society

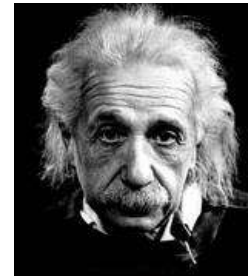
**Artificial Intelligence**

may either **exploit**

or **overcoming**

our **Natural Stupidity**

“Two things are infinite: the universe and human stupidity;  
and I'm not sure about the universe.”



SORRY for such a PESSIMISTIC TALK

not very funny

but

I wish you get the message of

**the Optimism the WILL** (Gramsci)

**and the Pleasure/Beauty of AI**

**END**

Thank you for your attention!

I like to thank our research group in Cognitive Science at ISTC:

the ‘*GOAL group*’

*Rino Falcone*

*(Emiliano Lorini)*

*Maria Miceli*

*Fabio Paglieri*

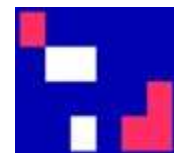
*Giovanni Pezzulo*

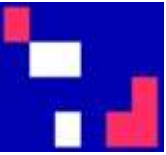
*Luca Tummolini*

.....

And IN MEMORY of *Rosaria Conte* (*Social Simulation LABSS Group*)

<http://www.istc.cnr.it/group/goal>





# Consequentialism

Giovanni Sartor



# The concept of consequentialism

- An action is morally required
  - iff it delivers that best outcome, relative to its alternative
  - Iff its good outcomes outweigh its negative outcomes to the largest extent
  - Iff it produces the highest utility?
- Morality as an optimisation problem!
- Various kinds of consequentialism
  - What are the good and bad things to be maximised?
  - How many there are?
  - How much each of them matters?
  - Can we construct a single utility function that combines gains and losses over multiple valuable goals?

# The reference approach: Utilitarianism

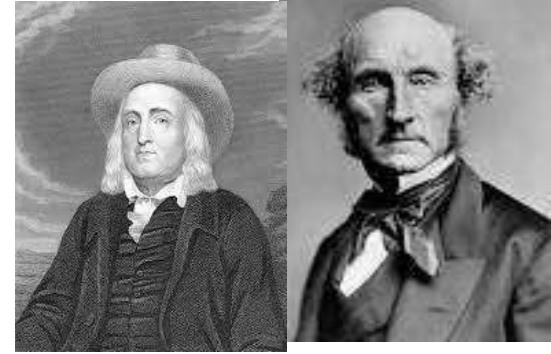
- Jeremy Bentham,
- John Stuart Mill. From Utilitarianism (1861). Principle of utility:
  - Actions are right in proportion as they tend to promote happiness, wrong as they tend to produce the reverse of happiness. By happiness is intended pleasure, and the absence of pain; by unhappiness, pain, and the privation of pleasure
- Utility: Happiness or satisfaction of desires/interests
- Utilitarianism is not egoism
  - The utility of everybody has to be taken into account equally



# Advantages of utilitarianism

- Conceptually simple
- Egalitarian (everybody's utility counts in the same way)
- Fits with some basic intuitions (making people happy is good, making them suffer is bad)
- In many case it is workable, in some cases problematic (what should we do about hunger, how shall we treat friends and relatives, etc.)





# Two versions of utilitarianism

- Act utilitarianism
  - Do the action that maximises utility
  - Do the optifimic action
- Rule utilitarianism
  - Follow the rule the consistent application of which maximises utility
  - Follow the optifimic rule
- Is AI utilitarian
  - What utility function would be utilitarian?
  - Should AI systems adopt an utilitarian reward function?
  - Should they go for the Act or the Rule versions (are they Archangels or Proles?)

# Issues with act utilitarianism

- Does it provide a good decision procedure
  - Can we choose what to do by optimising the outcome our actions? Do we have the information to make this calculation? Can an AI system have the information?
- Does it provide a good standard for assessing decisions?
- What is the link between utility and a reward function?

# Act utilitarianism: Problems

- Is it too demanding.
  - Should I give to the poor all that I have above the minimum that allows me to survive?
  - Should I give the same importance to everybody, regardless of their connection to me?
  - Is it OK to harm some people for the greater benefit of others
    - Reprisals? Torture? Sadism?
- What could an utilitarian say:
  - The cases in which utilitarianism seems to fail are not realistic
  - There is no real contrast between utilitarianism and mainstream moral beliefs

# Rule utilitarianism

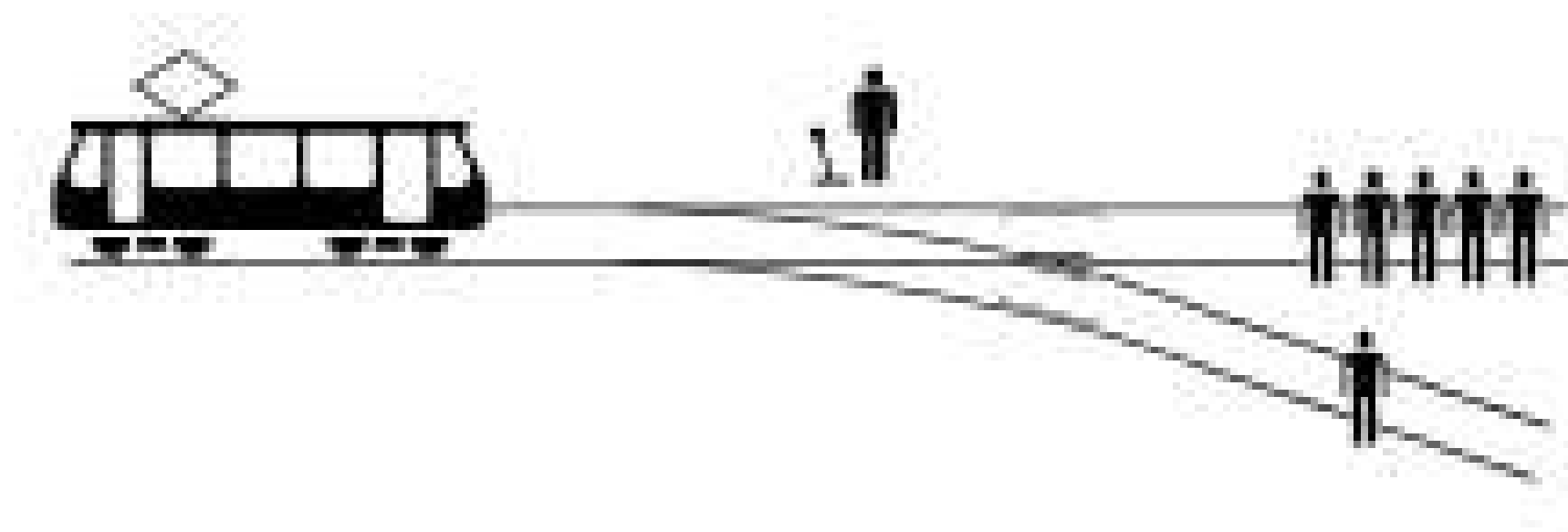
- an action is morally right just because it is required by an optimific social rule (a social rule the general compliance with which would provide the highest utility)
  - It is ok to tell the truth, not to steal, etc. since the general compliance with such norms would deliver the greatest utility
  - What about those exceptional cases in which the rule does not deliver
  - What is you know that most people are not following the rule.
    - Should we be honest if most people around as are dishonest?



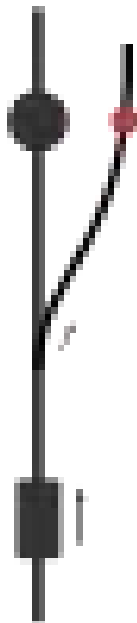
# A further issue: distribution

- Does it matter how the good and bad outcomes are distributed?
  - It is ok to make an action that benefits some to the detriment of others?
  - Always if the benefits outweigh disadvantages?
- Utilitarianism vs wealth maximisation
  - Utilitarianism favours (modest) redistribution of wealth, since the same amount of money gives more utility to the poor than to the rich
  - The impact of redistribution on wealth generation however has to be considered
- Wealth maximisation (adopted by some economic approach) aims at maximising the wealth in society regardless of distribution.

# The trolley problem



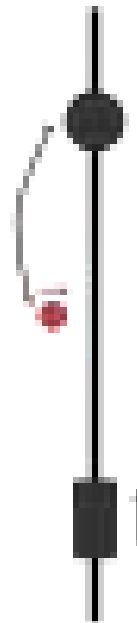
What would you do? What should an AI system tasked with monitoring traffic do



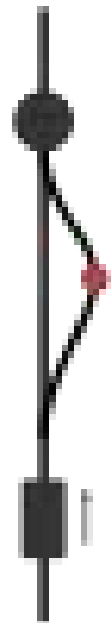
the switch  
Frost, 1947



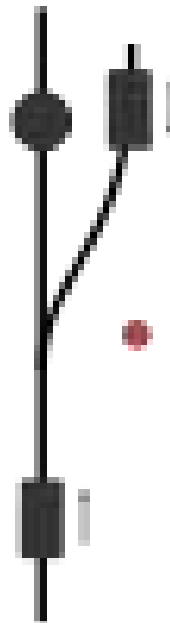
the fat man  
Brennan, 1976



the fat villain

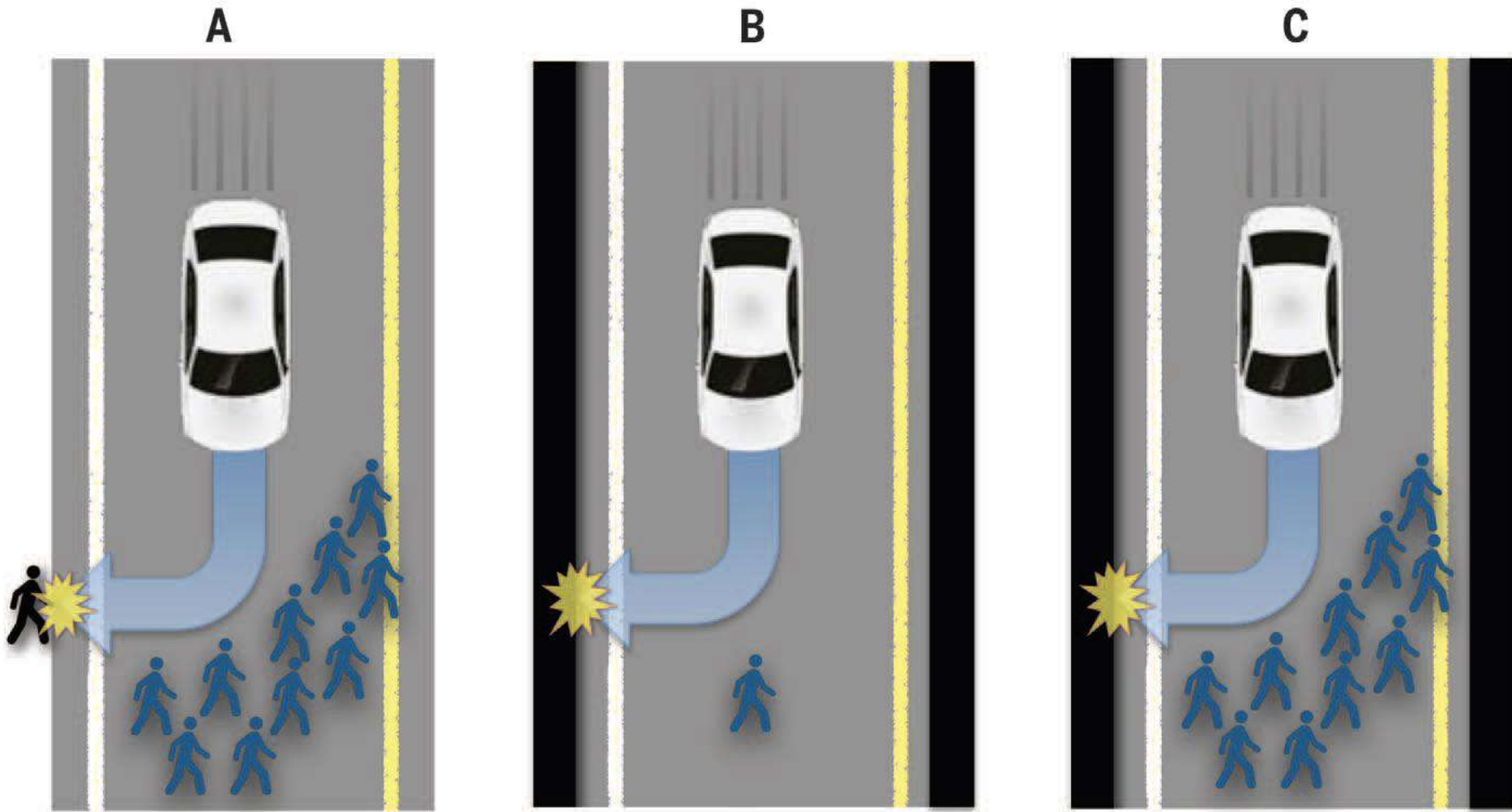


the loop  
Cassidy, 1968



the man in the yard  
Koger, 1992

# The social dilemma of autonomous vehicles

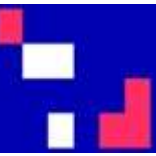




# Judith Jarvis Thomson: The surgeon case



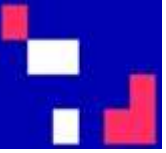
- A brilliant transplant surgeon has five patients, each in need of a different organ, each of whom will die without that organ. Unfortunately, no organs are available to perform any of these five transplant operations.
- A healthy young traveler, just passing through the city in which the doctor works, comes in for a routine checkup. In the course of doing the checkup, the doctor discovers that his organs are compatible with all five of his dying patients.
- Suppose further that if the young man were to disappear, no one would suspect the doctor. 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?



# Thanks for your attention!

[giovanni.sartor@unibo.it](mailto:giovanni.sartor@unibo.it)





# Ethics/Morality

Giovanni Sartor



# What is morality/ethics

- In deciding what to do, or in evaluating what other do:
  - We can take our individual perspective, focusing on our particular interests (self-interest) or
  - We can be motivated by the belief that an action is right, regardless of how it affect our interest (morality/ethics)
- Positive (conventional) morality: the moral rules and principles that are accepted in a society
  - Can there be bad positive morality?
- Critical morality
  - The morality that is correct, rational, just (maybe since considers all individual and social interests at stake giving each one the due significance (harms to other, impacts on environment, etc.)
- We can criticise positive morality based on our critical morality: we may be right or wrong (e.g., feminist critiques against patriarchy, nazi criticism against being compassionate)

# What is morality/ethics

- In deciding what to do, or in evaluating what other do:
  - We can take our individual perspective, focusing on our particular interests (self-interest)
  - We can be motivated by the belief that an action is right, regardless of how it affect our interest (morality)
- Positive (conventional) morality: the moral rules and principles that are accepted in a society
  - Can there be bad positive morality
- Critical morality
  - The morality that you believe is correct, rational, just (maybe since considers all individual and social interests at stake giving each one the due significance (harms to other, impacts on environment, etc.)
- We can criticise positive morality based on our critical morality:
  - We may be right or wrong in our criticism (e.g., feminist critiques against patriarchy, nazi criticism compassion and universalism, etc.)

# Ethics vs metaethics

- Normative ethics is concerned with determining what is morally required, how one ought to behave
- Metaethics is concerned with is the study of the nature, scope, and meaning of moral judgement
  - Can ethical judgments be true or false?
    - What is the difference between
      - I prefer vegetables to meat
      - I ought to eat more vegetables to be more healthy
      - We ought to become vegetarians
  - Do they correspond to some facts in the world?
    - What facts make it true that we ought to become vegetarian? Or that we ought not to harm others?
  - Does ethic pertain to rationality of or to feelings
    - David Hume: is not contrary to reason to prefer the destruction of the whole world to the scratching of my **finger**. Morality is a matter of sentiment (of impartial spectators)
    - Emmanuel Kant: we can know what is moral through our reason
    - David Ross: we can know what is more through our intuition

# Absolutism vs Relativism

- There is a single true ethics: when two people express incompatible ethical judgement one of them must be wrong
- Ethical judgements are always relative to particular frameworks of attitudes
  - A statement such as “abortion is morally permissible” or “adultery is prohibited” or “killing a willing person is wrong” may be true under some framework and wrong under some other
  - An analogy: as in mechanics, judgements are relative to the frame of reference (a body may be moving relative to one frame and stationary relatively to another), so is in morality

# Morality and disagreement

- Morality is a place for widespread disagreement
  - Abortion
  - Migration
  - Capital punishment
  - Humanitarian wars
  - ...
- But there is something on which we may agree?
  - It is wrong to kill innocent people?
  - It is (usually) wrong to lie?
  - It is (usually) wrong to harm people?

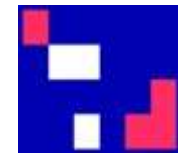


# Pro-tanto and all-things-considered moral judgement

- Many moral prescriptions are defeasible. They state general propositions that are susceptible of exceptions.
  - We should not lie
  - What if a lie would save a person's life?
- Do we want a robotic agent to take its duties as defeasible?
- An act is a ***prima facie* duty** when there is a moral reason in favor of doing the act, but one that can be outweighed by other (moral) reasons.
- David Ross: "If I have promised to meet a friend at a particular time for some trivial purpose, I should certainly think myself justified in breaking my engagement if by doing so I could prevent a serious accident or bring relief to the victims of one."

# Morality and other normative systems

- Law
  - Does positive or critical morality include all laws enforced by the state? Does it include only such laws?
- Religion
  - Does critical morality include all and only what has been commanded by God
  - Did God command something because it was moral, or did anything become moral for having been commanded by God (rationalism vs voluntarism). What about Abraham and Isaac.
  - Are atheists necessarily immoral or amoral? Is an atheistic society necessarily more immoral than a religious society?
- Tradition
- Self interest:
  - may morality and self interest collapse: should we do all and only what fits our personal interest (Gige's ring)



# Thanks for your attention!

[giovanni.sartor@unibo.it](mailto:giovanni.sartor@unibo.it)





**POLITECNICO**  
MILANO 1863

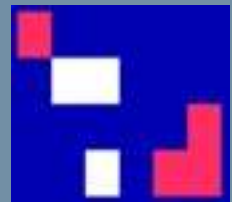
**MAI4CAREU**

Master programmes in Artificial  
Intelligence 4 Careers in Europe

# Do Artifacts Have Politics?

**Viola Schiaffonati**

*Artificial Intelligence and Robotics Lab*  
*Dipartimento di Elettronica, Informazione e Bioingegneria*





# Robert Moses's overpasses



# Racists overpasses

- *Robert Moses (1888-1981) was a very influential and contested **urban planner***
- *He designed several **overpasses** over the parkways of Long Island which **were too low to accommodate buses***
- *Only cars could pass below them and for that reason the overpasses complicated access to Jones Beach Island*
- ***Only people who could afford a car** – and in Moses' days there were generally not Afro-Americans – could easily **access the beaches***



# “Do artifacts have politics?”

*“Robert Moses, the master builder of roads, parks, bridges, and other public works from the 1920s to the 1970s in New York, had these overpasses built to specifications that would **discourage** the **presence of buses** on his **parkways**. According to evidence provided by Robert A. Caro in his biography of Moses, the reasons reflect **Moses's social-class bias** and **racial prejudice**. Automobile owning whites of “upper” and “comfortable middle” classes, as he called them, would be free to use the parkways for recreation and commuting. **Poor people** and **blacks**, who normally used public transit, **were kept off the roads** because the **twelve-foot tall buses** could **not** get through the **overpasses**. One consequence was to **limit access** of **racial minorities** and **low-income groups** to Jones Beach, Moses's widely acclaimed public park.”*

(Winner 1980)

# Agenda

- Technological artifacts as morally and politically charged
  - **Technological mediation**
  - The **moralization** of **technologies**
- From passive to **active responsibility**
- AI technologies
  - **Experimental** technologies
  - The **invisibility factor**
- **Criticizing** the moral character
- **Ethics** of **engineering design**

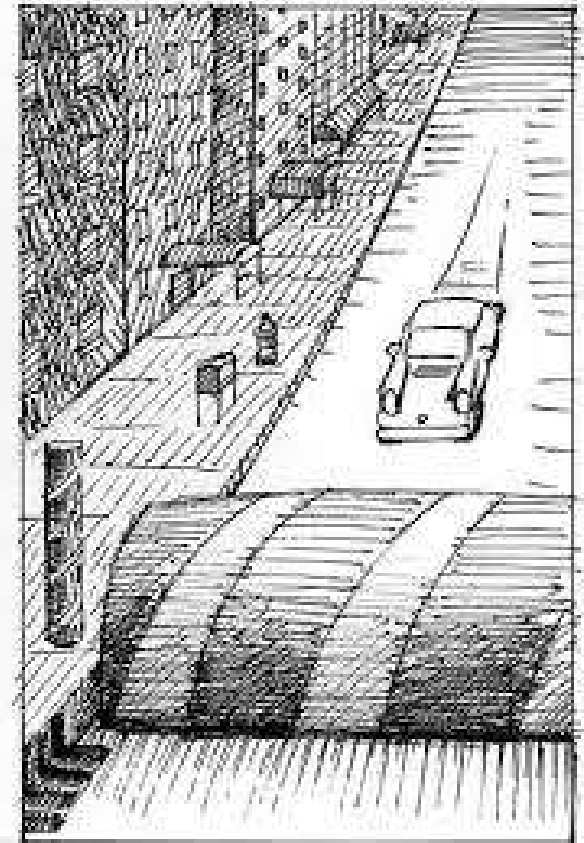


# Beyond racist overpasses

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

# Ethics as a matter of things

- **Artefacts** are bearers of **morality**, as they are constantly taking all kinds of moral decisions for people (Latour 1992)
  - Ex.: moral decision of how fast one drives is often delegated to a speed bump which tells the driver "*slow down before reaching me*"



# Technological mediation



- The phenomenon that when technologies fulfill their functions, they also help to **shape actions** and **perceptions** of **their users**
- Technologies are **not neutral “intermediaries”** that simply connect users with their environment
- They are **impactful mediators** that help to shape how people use technologies, how they experience the world and what they do

# Mediation of perception: obstetric ultrasound

- Ultrasound is not simply a **functional means** to make visible an unborn child in the womb, but **mediates** the relations between the fetus and the parents

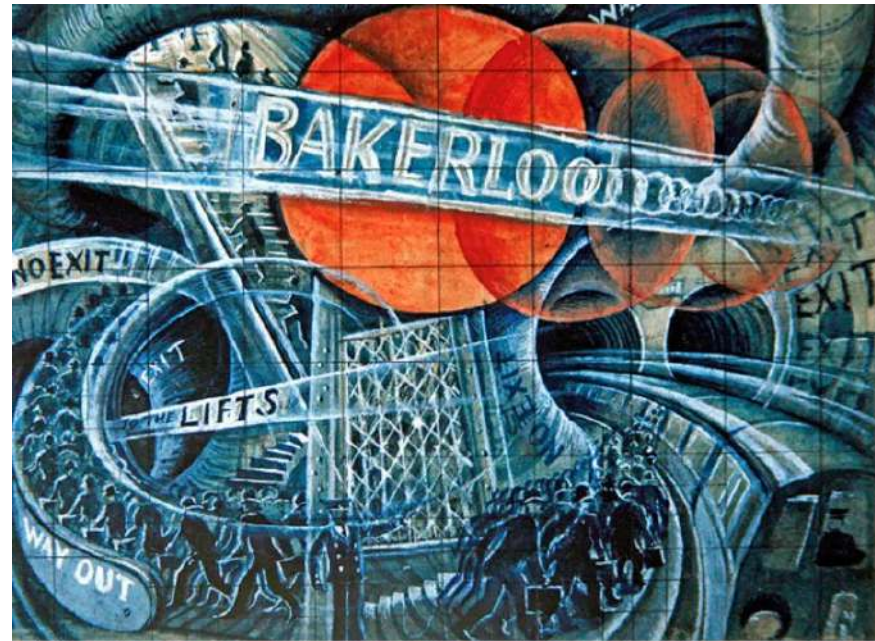


# Obstetric ultrasound and translations

- Number of **translations** of the relations between expecting parents and the fetus while mediating their visual contact
  - Ultrasound isolates the fetus from the female body: **new ontological status of the fetus** as a separate living being
  - Ultrasound places the fetus in a context of medical norms: it translates **pregnancy into a medical process**, the fetus into a possible patient, and congenital defects into preventable sufferings (**pregnancy as a process of choices**)
- **Ambivalent role** of ultrasound: it may both encourage abortion (prevent suffering) and discourage it (emotional bonds)

# Moralizing technologies

- Instead of moralizing other people humans should/could also **moralize their material environment**
  - 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 responsibility ...**
- **Responsibility** is connected to being held **accountable** for your **actions** and for the **effects** of your actions
  - Making of choices, taking decisions, failing to act, ...
- **Passive** responsibility is a **backward-looking** responsibility which is relevant **after** something **undesirable occurred**



# ... to active responsibility



- **Active responsibility** means **preventing** the **negative effects** of technology but also **realizing** certain **positive effects** (Bovens 1998)
- **Value sensitive design:** **moral considerations** and values are used as **requirements for the design** of technologies (Friedman 1996, van der Hoven 2007)



# Active responsibility and AI

*"I will call technologies **experimental** if there is only **limited operational experience** with them, so that social benefits and risks cannot, or at least not straightforwardly, be assessed on basis of experience."*

(van de Poel 2016)

- **Uncertainty** that is inherent in the **introduction** of these new technologies (sophisticated **AI** systems) into **society**



# AI and 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.”

(Moor 1985)

# Types of invisibility

- Invisibility of **abuse**

*"Invisible abuse is the intentional use of **invisible operations** of a computer to engage in **unethical conduct**. A classic example is the case of a programmer who realized he could steal excess interest from a bank."*

- Invisibility of **programming values**

*"Consider for example computerized airline reservations. Many different programs could be written to produce a reservation service. American Airlines once promoted such a service called SABRE. This **program** had a **bias** for American Airline flights built in so that sometimes an American Airline flight was **suggested by the computer** even if it **was not the best flight** available."*

- Invisibility of **complex calculations**

*"Computers today are capable of **enormous calculations beyond human comprehension**. Even if a program is understood, it does not follow that the calculations based on that program are understood."*

# Moralizing technologies (Verbeeck 2011)

- Many of our **actions** and **interpretations** of the world (also moral ones!) are **co-shaped by the technologies**
- **Moral decision-making** is a **joint effort** of **human beings** and **technological artefacts**



<https://www.youtube.com/watch?v=S8a1DascnZg>

# Taking mediations into ethics

- **Alcohol lock for car**  
(car lock that analyzes your breath)
- **Smart showerhead**  
(showerhead that regulates and reduces the flux of water to save water)





# Alcohol lock for cars



- **Alcohol lock for car** (car lock that analyzes your breath): “*Don’t drive drunk*”
- Suppose that a car with such a system is not more expensive than the one without it and works perfectly

*How many of you would buy such a car? Why?*

*How many of you would not buy such a car? Why?*

# Taking mediations into ethics



- **Smart showerhead**  
(showerhead that regulates and reduces the flux of water to save water): *"Don't waste water"*
- Suppose that this showerhead is not expensive and allows you to save 50% of your daily consumption of water

*How many of you would buy it?  
Why?*

*How many of you would not  
buy it? Why?*

# Criticizing the moral character

- Variety of **negative reactions** to explicitly **behavior-steering technologies** (also when they are for the good!)



- Fear that **human freedom** is threatened and that democracy is exchanged for **technocracy**
  - **Reduction of autonomy** perceived as a threat to **dignity**
  - Not humans but **technologies** are in **control**
- Risk of **immorality** or **amorality**
  - Form of **moral laziness** with behavior-steering technologies



# A democratic way to moralize technology?

- **Technologies** differ from **laws** in **limiting human freedom** because they are not the result of a democratic process
  - See the difference between the alcohol lock for car and the smart showerhead
- It is important to find a **democratic way** to “**moralize technology**”
  - The processes used to insert values must be transparent and **publicly discussed**



# Designing mediations

- Designers cannot simply “inscribe” a desired form of morality into an artefact
- In order to build in specific forms of mediation in technologies, designers need **to anticipate the future mediating role** of the **technologies** they are designing
  - **Unintentional** and **unexpected forms of mediation** (ex.: energy-saving light bulbs used in places previously left unlit and hence increasing energy consumption)



# Not only desired forms

- Designers cannot simply “inscribe” a desired form of morality into an artefact, because this also depends on
  - **Users** that interpret technologies
  - **Technologies** themselves which can evoke **emergent** forms of mediation

# Strategies for designing mediations

- **Anticipating mediation by imagination**
  - Trying to imagine the ways technology-in-design could be used to deliberately shape user operations and interpretations
- Augmenting the existing design methodology of **Constructive Technology Assessment (CTA)**
  - **CTA** is an approach in which TA-like efforts are carried out **parallel to the process of technological development** and are **fed back** to the development and design process
  - Not only to determine what a technology will look like, but all **relevant social actors**

# Ethics of engineering design

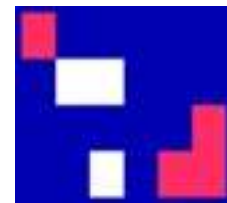
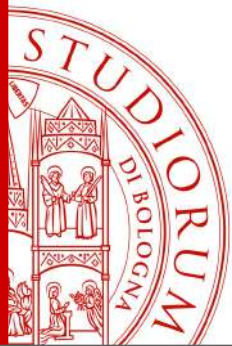
- **Technology design** appears to entail **more than inventing functional products**
- The perspective of technological mediation reveals that **designing** should be regarded as a **form of materializing morality**
- The **ethics of engineering design** should take more seriously the **moral charge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

# References

- Latour, B. (1992). "Where Are the Missing Masses? The Sociology of a Few Mundane Artifacts" in Wiebe E. Bijker and John Law, eds., *Shaping Technology/Building Society: Studies in Sociotechnical Change*, Cambridge, Mass.: MIT Press, 1992, pp. 225–258
- Van de Poel, I. and Royakkers, L. (2011). *Ethics, Technology, and Engineering*, Wiley-Blackwell
- Verbeek, P. P. (2011). *Moralizing Technologies*, University of Chicago Press.
- Winner, L. (1980). "Do artifacts have politics?", *Daedalus*, 109, 121-136

**MAI4CAREU**

Master programmes in Artificial  
Intelligence 4 Careers in Europe



# Responsibility and automation in Socio- technical systems

*The case of air traffic management*

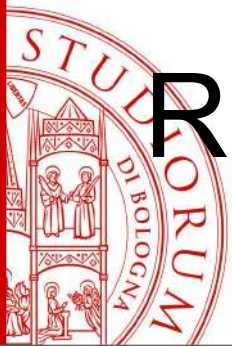
Giuseppe Contissa



Co-financed by the European Union  
Connecting Europe Facility

This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423





# Responsibility and automation

- 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?





# “responsibility”

As captain of the ship, X was **responsible** for the safety of his passengers and crew. But on his last voyage he got drunk every night and was **responsible** for the loss of the ship with all aboard.

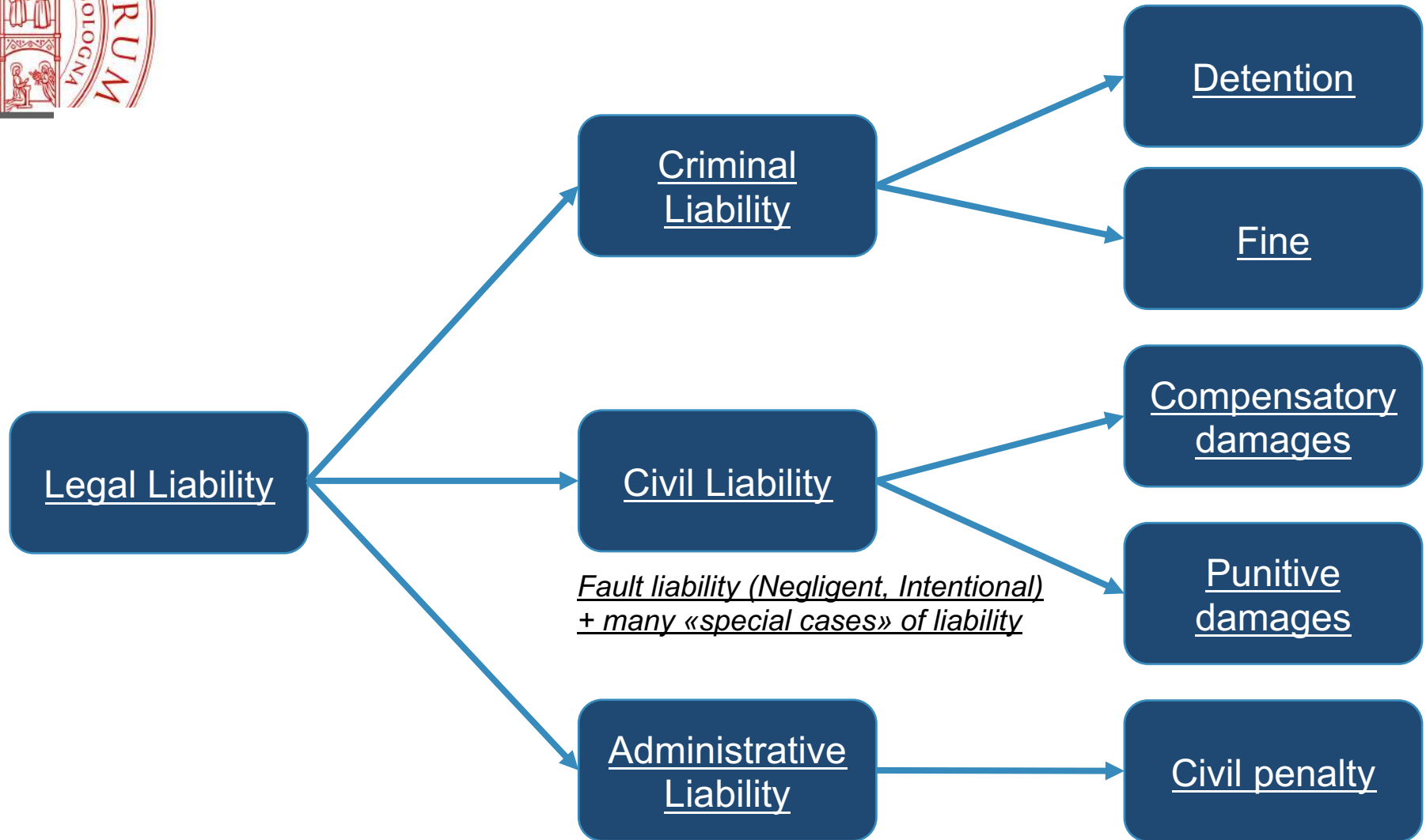
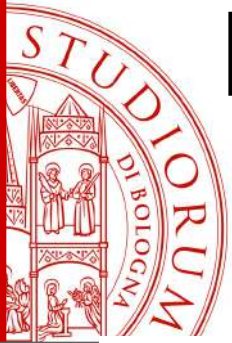
It was rumoured that he was insane, but the doctors considered that he was **responsible** for his actions. Through out the voyage he behaved quite **irresponsibly**, and various incidents in his career showed that he was not a **responsible** person.

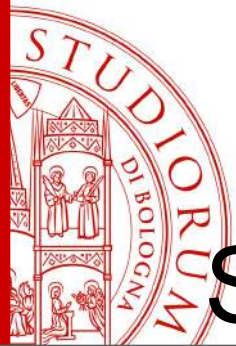
He always maintained that the exceptional winter storms were **responsible** for the loss of the ship, but in the legal proceedings brought against him he was found criminally **responsible** for his negligent conduct, and in separate civil proceedings he was held legally **responsible** for the loss of life and property.

He is still alive and he is morally **responsible** for the deaths of many women and children.

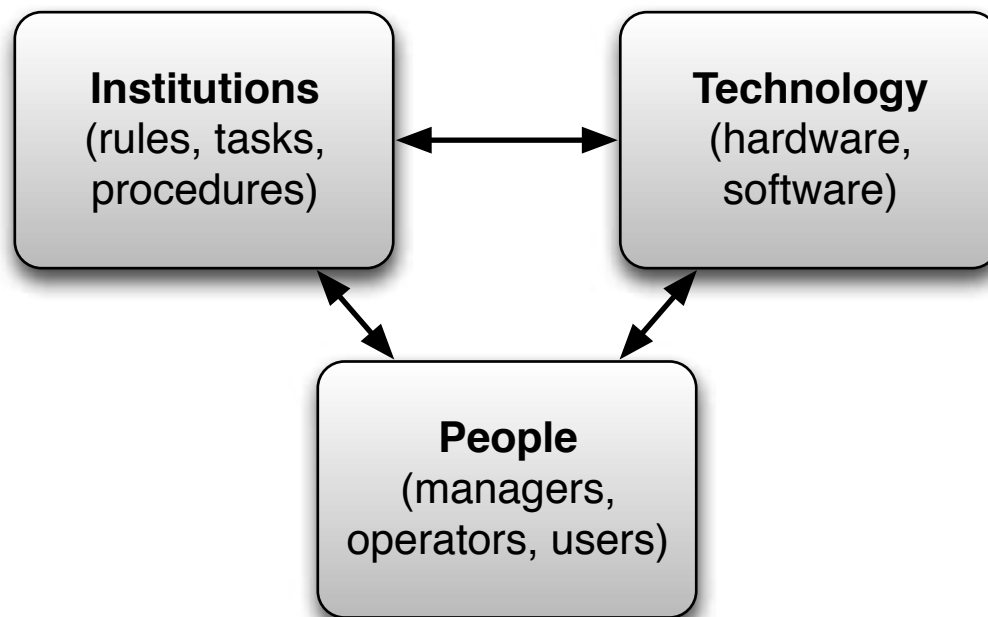
*(Hart, H.L.A., Punishment and Responsibility: Essays in the Philosophy of Law, 1970)*

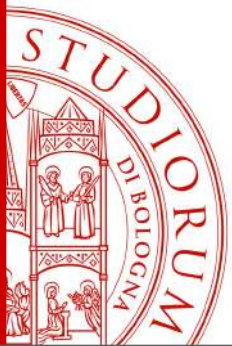
# Liability (legal responsibility)





# Socio-technical systems: basic structure

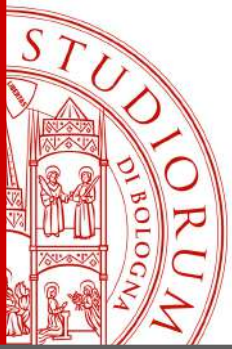




# Socio-technical systems: examples







# The future of ATM

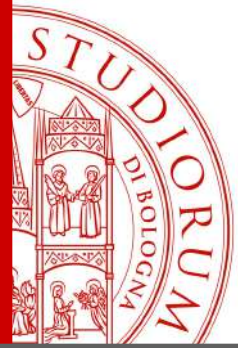
- In the time horizon of SESAR, that is over the next 30 years, a new generation of air traffic management systems will be developed.
- Such systems will be highly automated. They will make choices and engage in actions with some level of human supervision, or even without any such supervision.



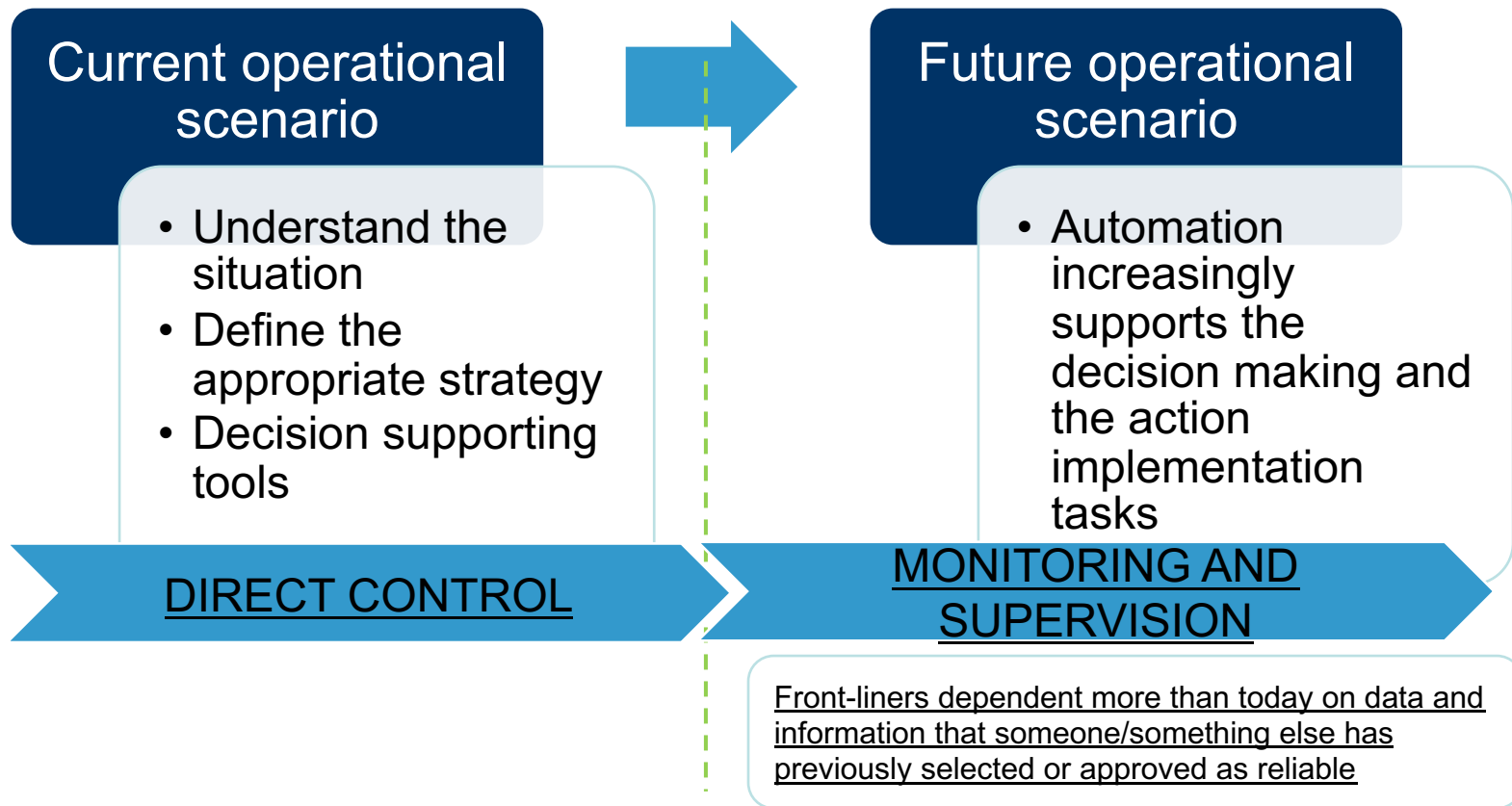
# Automation and the future ATM scenario



- New generation of ATM systems to increase capacity, safety, efficiency and sustainability
- Higher levels of automation



# AUTOMATION SUPPORT







# Implications of automation

- Delegation of task from operators to technology
- Humans as controllers and supervisors
- Hybrid agency (symbiosis/coagency → joint cognitive systems)
- Machine intelligence and autonomy (= independence + cognitive skills)
- The challenge of complexity (technological, “many hands”)





# Automation: not all or nothing

- Not just **substitution of a human operator**
- Support to human capabilities in performing tasks

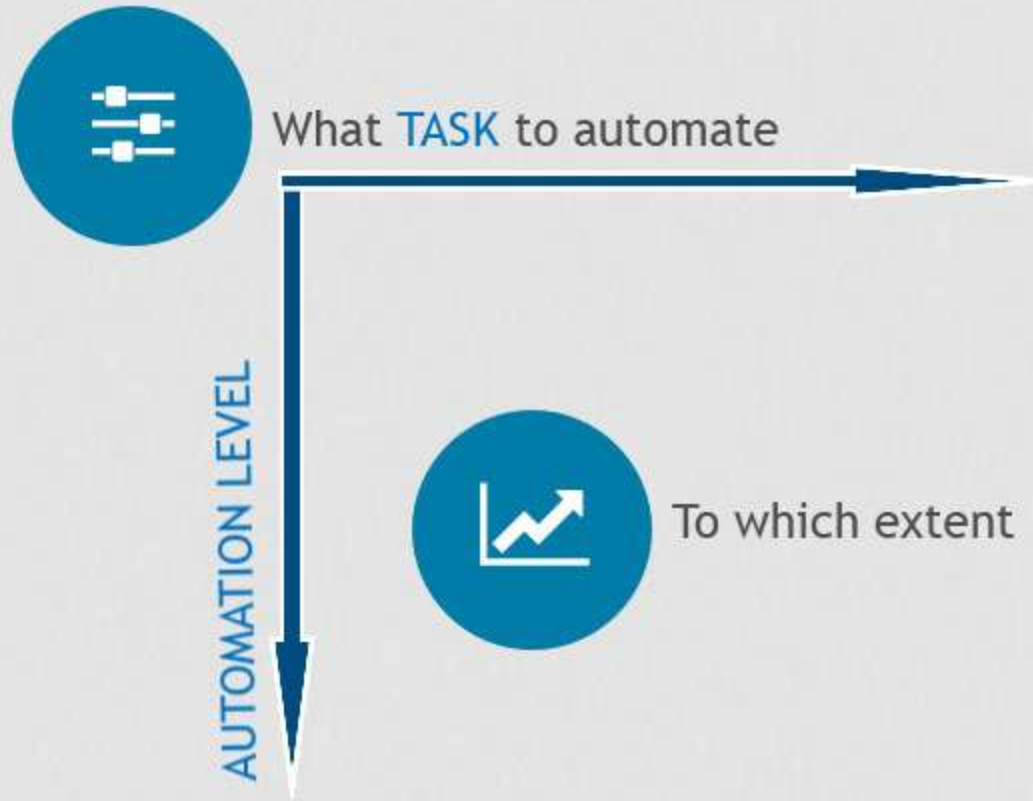


- Some degree of cooperation is usually required



# Automation: not all the same

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





# The level of automation taxonomy (SESAR 1)

From INFORMATION to ACTION

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

A condensed version of the LOAT matrix

# ROT / Use of video cameras in the control tower

## **A** **INFORMATION** **ACQUISITION**

**A0** Manual Information Acquisition

**A1** Artefact Supported Information Acquisition

**A2** Low Level Automation Support of Info Acquisition

**A3** Med. Level Automation Support of Info Acquisition

**A4** High Level Automation Support of Info Acquisition

**A5** Full Automation Support of Info Acquisition



The system supports the human in acquiring information on the process s/he is following. Filtering and/or highlighting of the most relevant information are up to the human.



# Activation of speed vectors by controllers

## **B** **INFORMATION** **ANALYSIS**

**B0** Working-memory based  
Information Analysis

**B1** Artefact Supported  
Information Analysis

**B2** Low Level Automation  
Support of Info Analysis

**B3** Med. Level Automation  
Support of Info Analysis

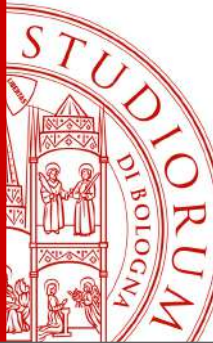
**B4** High Level Automation  
Support of Info Analysis

**B5** Full Automation  
Support of Info Analysis



**Based on user's request**, the system **helps** the human in comparing, combining and analysing different information items regarding the status of the process being followed.

# AMAN sequence of landing aircraft

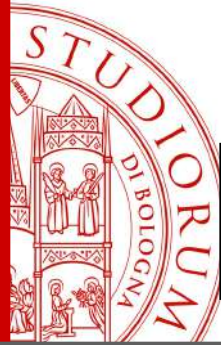


<b>C</b> <b>DECISION AND ACTION SELECTION</b>	
<b>C0</b>	<u>Human Decision Making</u>
<b>C1</b>	<u>Artefact Supported Decision Making</u>
<b>C2</b>	<u>Automated Decision Support</u>
<b>C3</b>	<u>Rigid Automated Decision Support</u>
<b>C4</b>	<u>Low Level Automatic Decision Making</u>
<b>C5</b>	<u>High Level Automatic Decision Making</u>
<b>C6</b>	<u>Full Automatic Decision Making</u>

The screenshot displays a software interface for AMAN. At the top, there are tabs for 'Config', 'NonSeq', 'Meteo', and 'TLM'. Below the tabs, a header row shows 'S01R 01L:3.0 01R:3.0 19L:3.0 19R:3.0 01L/01R:3.0 19L/19R:3.0'. The main area is a grid with time on the vertical axis (from 15:20 down to 14:24) and aircraft data on the horizontal axis. The data includes aircraft identifiers (e.g., 01R SUM SAG005, 01R TOR SNB612), aircraft types (e.g., BE20 L, B738 M), and numerical values (e.g., 127, 151). On the right side, there are additional columns with labels like 'SIG 01R CN' and 'HES/SIG'. At the bottom, there are buttons for 'I', 'W', 'E', 'TRK', 'FPL', 'MET', and a page number '14 28'. A URL is provided at the bottom of the screenshot: <https://www.eurocontrol.int/sites/default/files/article/content/documents/nm/fasti-aman-status-review-2010.pdf>, page 16.

The system proposes one or more decision alternatives to the human, leaving freedom to the human to generate alternative options. The human can select one of the alternatives proposed by the system or her/his own one.

# Autopilot



## D ACTION IMPLEMENTATION

**D0** Manual Action and Control

**D1** Artefact Supported Action Implementation

**D2** Step by step Action Support

**D3** Low Level Support of Action Sequence Execut.

**D4** High Level Support of Action Sequence Execut.

**D5** Low Level Automation of Action Sequence Exec

**D6** Medium Level Automat. of Action Seq. Execut.

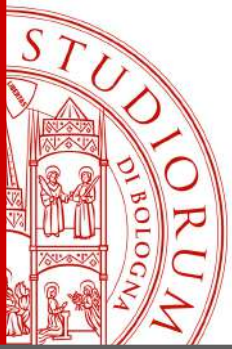
**D7** High Level Automation of Action Seq. Execut.

**D8** Full Automation of Action Sequence Exec



The system performs automatically a sequence of actions after activation by the human. The human can monitor all the sequence and can interrupt it during its execution.





# Some questions

---

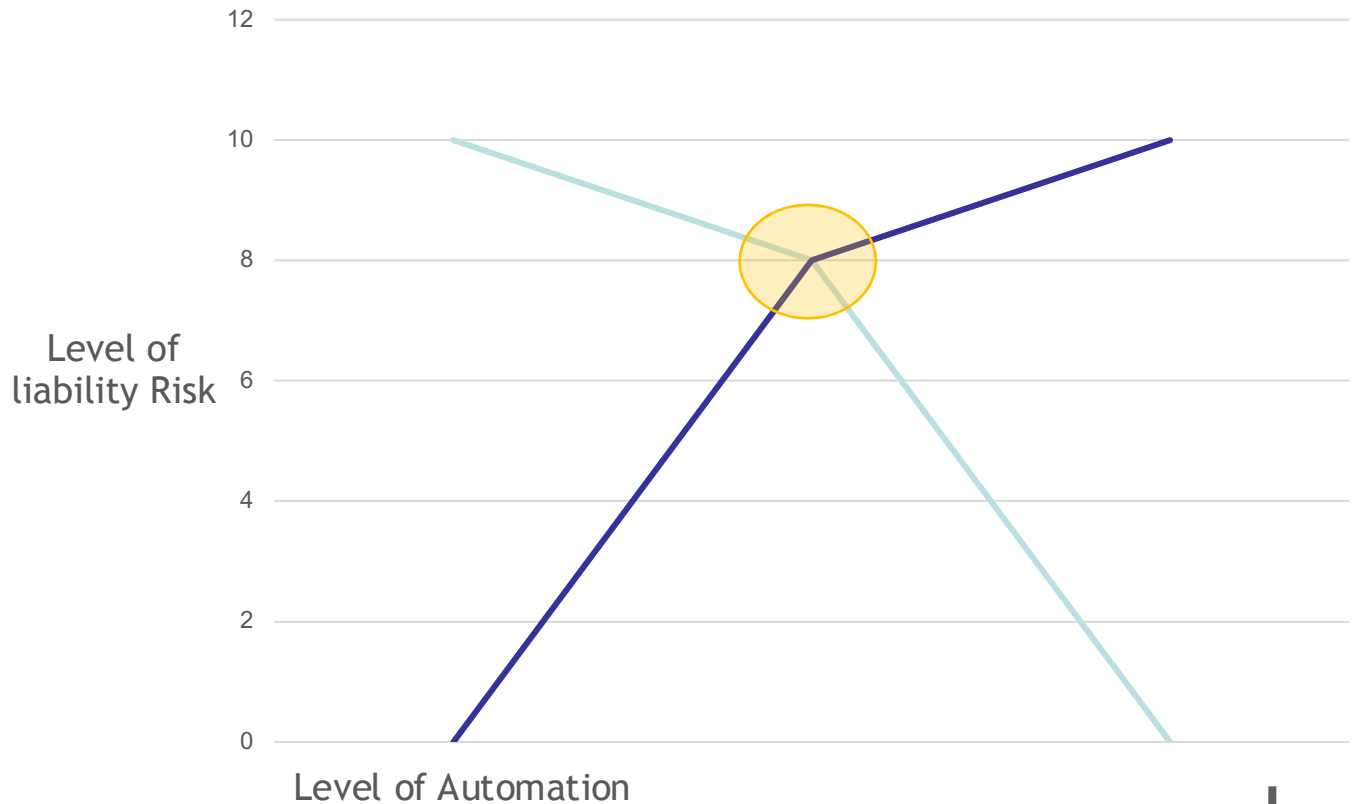
- How automation transforms operators' roles and tasks? What impact on their responsibilities?
- Who is responsible for the behaviour of systems that humans cannot fully monitor and control?
- Who is responsible for information supplied by automated systems that the human cannot verify?



# Level of automation and liability risk



- Increasing the level of automation will proportionally increase the liability risk for the **technology provider** and decrease the liability risks for the **human operator**.
- However, the employment of technologies with **intermediate levels of automation** may result in a high liability risk both for the technology provider and the human operator

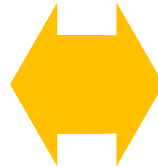


# Fragmentation of tasks and liability

The **fragmentation of tasks** may results in uncertainty and complexity of procedures

## Human operator

- difficult to asses how and who should carry out each task
- high liability risk for negligence



## Technology provider

- difficult to design HMI to adequately support decision making and/or to provide exhaustive information
- high product liability risk, caused by design and information defects.



# Highly automated systems/AI systems: liability shift

**Liability** for injury/harm caused by technological failure **gradually** transferred to the **organisation(s) developing / using/ maintaining** the technology

Grounds for the attribution:

- Product liability (no-fault liability, grounded on **defectiveness**, in particular wrt design defects and warning defects)
  - Organisational / no-fault liability: generation of risks and ability to prevent them (and possibility to distribute losses)
  - Vicarious liability (for faults of employees, residual)
  - In the future: Liability for failing to deploy automated/AI systems?
- 
- Liability assessment should be carried out as soon as possible in the life cycle of technology.
  - Liability allocation related to level of automation of technology, in particular to the cognitive functions of the automated/AI technology and on the H-M interaction
  - To be assessed in relation to role of technology in accidents



# Highly automated systems/AI systems: liability shift /2

## **Individual liability** (*criminal/civil*, fault liability) would persist

- only when the human acted with an intention to cause harm or with recklessness (e.g. Just Culture)?... Or..
- always, human as «moral crumple zone» (Elish 2018)?
- What about decisions taken by humans when interacting with automated/AI systems?



# Other important issues on liability

## – Liability and standards/certification

- Liability shield for the producer?
- “Legitimate” expectation for the user/operator?
- Liability of certicators / standard setters

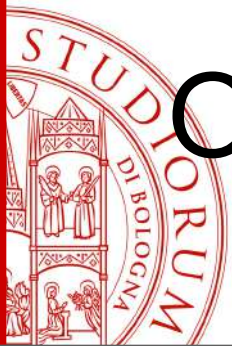
## – Right of recourse. Who will pay in the end?

- In complex systems, the *law may channel liability* towards one actor (e.g. in ATM, the air carrier), *but recourse against the one who had control* over the malfunctioning component of the system

## – The role of insurance

- Mandatory insurance for producers/manufacturers?
- Specific issues of highly automated /AI systems (cyber risk, wilful misconduct)

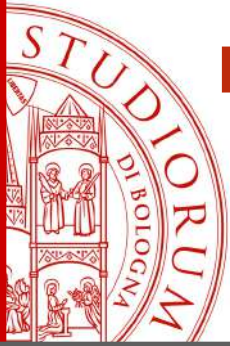
## – International context: “forum shopping”



# Open issue: Decision making authority

- **Effective decision-making authority in socio-technical systems**
  - Joint cognitive systems?
  - The model described (or prescribed) by laws, regulations, procedures:
    - Right not to be subject to (fully) automated individual decision-making (Art 22 **GDPR**): “[oversight of the decision] should be carried out by someone who has the authority and competence to change the decision” (Art29WP)
    - **Aviation**: ICAO Annex 2, sec. 2.3.1 Responsibility of pilot-in-command (ultimate authority, ultimate responsibility)
    - Vienna Convention on **Road Traffic**, Art. 1(v) "Driver" means any person who drives a motor vehicle or other vehicle (but amendments for ADS)
    - Art 14 new **AI ACT** proposal, human oversight for high-risk AI systems

# EFFECTIVE DECISION-MAKING AUTHORITY



What about decisions to be taken jointly with AI, in conditions of limited resources – time, information, explanations? E.g.:

- **Medical diagnosis** assisted by AI (Lagioia, Contissa 2020)
- **Frontex border** controls: «12 seconds to decide»



*Machine intelligence is fundamentally **alien**, and often, the entire purpose of an AI system is to learn to do or see things in ways humans cannot[..]*

*Ultimately, the **lack of a principled basis to contradict AI predictions implies that the reasonableness of an action in individual cases must be tied to the decision to use AI as a general matter.** (Selbst 2019)*

*Owing to the **evidence** in their favor (stipulated by definition), it is more appropriate to think of **expert robots as above average in their ability to make decisions that will produce desirable outcomes [...]***

*This fact suggests that **granting a general decision-making authority to human experts will be problematic once expert robots are properly on the scene.** (Millar, Kerr 2018)*



# Legal Case

Step 1  
Understand  
the Context

## GATE 1

check completeness and suitability of background information



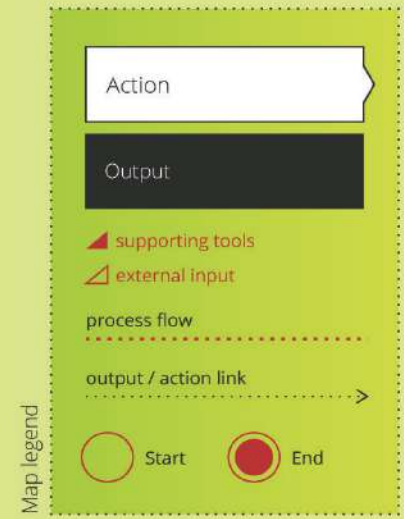
Step 2  
Identify  
Liability Issues



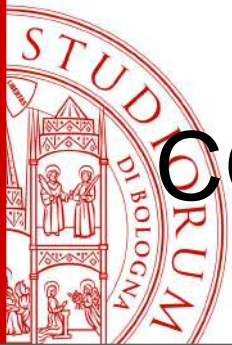
Step 3  
Address the  
Liability Allocation



Step 4  
Collect Finding and  
Systemic Analysis

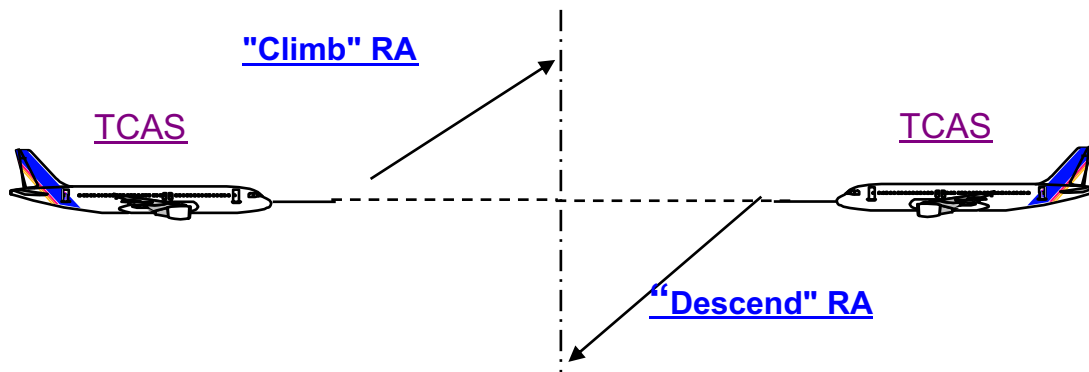




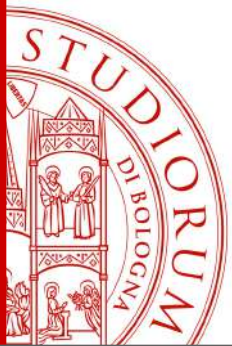


# ACAS/TCAS II (TRAFFIC COLLISION AVOIDANCE SYSTEM)

<http://www.skybrary.aero/index.php/TCAS>



- Visual and aural advices
- 2 types of advisories: **TA (Traffic Advisory)** and **RA (Resolution Advisory)**
- RA shall be executed by the crew; The system decides the best option and informs the human
- During the execution by the pilot the system provides guidance through continuous visual and aural feedback

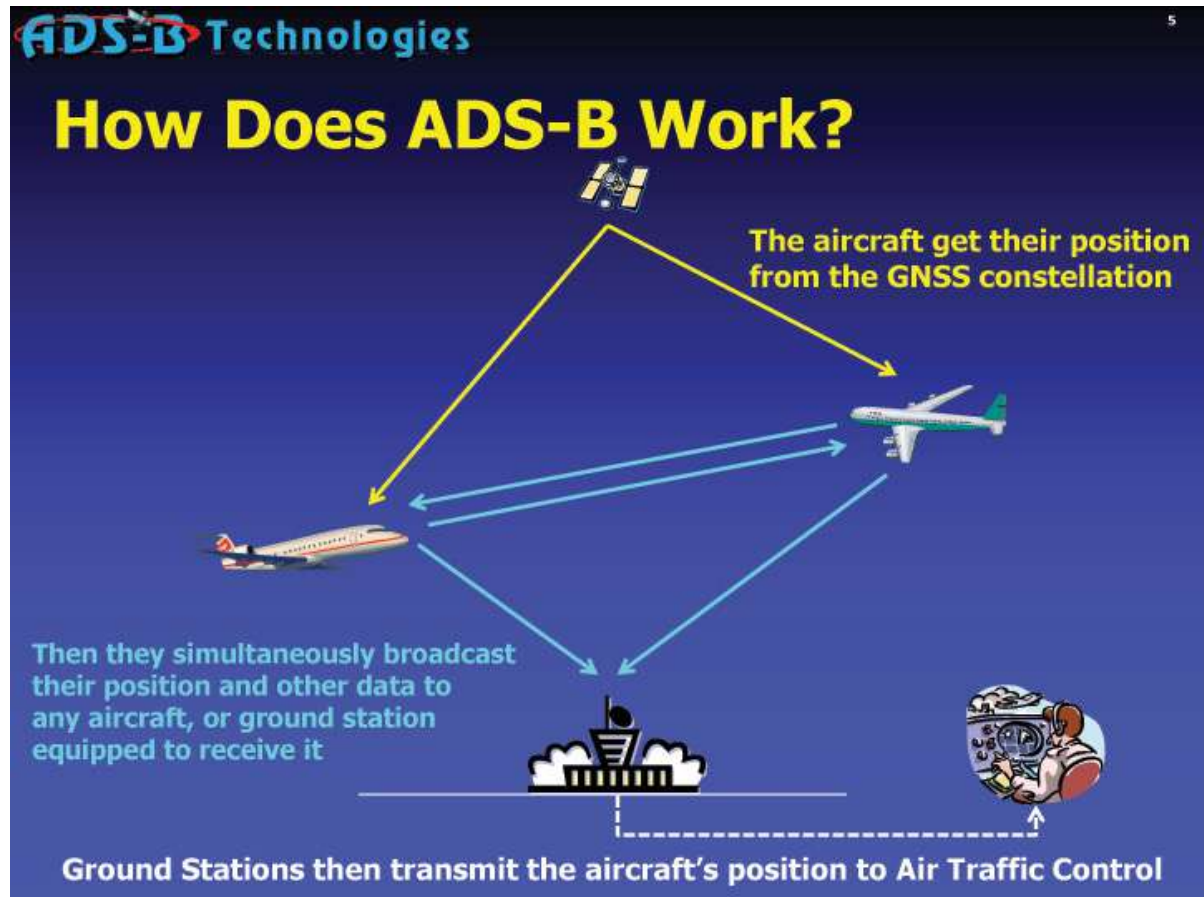


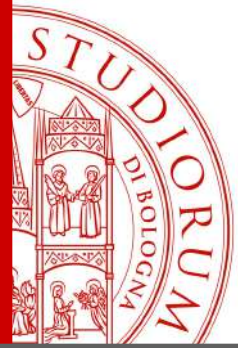
# ACAS X and ADS-B

ACAS X will replace the current generation of systems ACAS/TCAS II

It will use new sources of surveillance data, including ADS-B (Automatic Dependent Surveillance Broadcast)

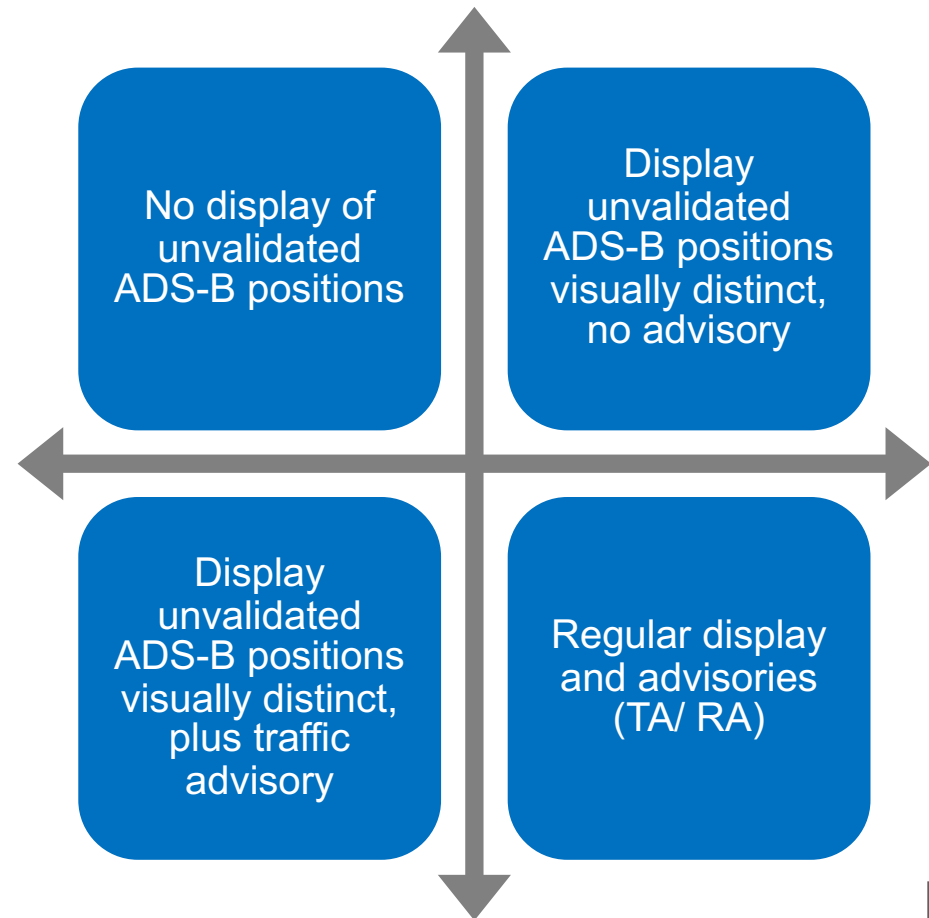
ADS-B is an enabler for the change from radar based towards satellite based aircraft location systems.





# FOCUS OF THE CASE STUDY: ACAS X AND UNVALIDATED ADS-B POSITIONS

- The treatment of **unvalidated ADS-B positions** by ACAS X emerged as one of the controversial design issues with respect to liability.
- ‘Unvalidated’ refers to positions which are solely based on ADS-B data, not validated through other surveillance data sources.
- 4 design options debated by EUROCAE as in the diagram.



# ARGOS V0.1







# ARGOS modes of operations

L3

## ARGOS AS A DECISION SUPPORT TOOL

For all flights, ARGOS displays the best plan. The ATCO can approve the plan, impose a constraint to let ARGOS revise the plan, or come up with his/her own plan. For CPDLC flights, ARGOS executes the plan. For non-CPDLC flights, the ATCO is reminded and the plan is the default selection in the menus.

L5

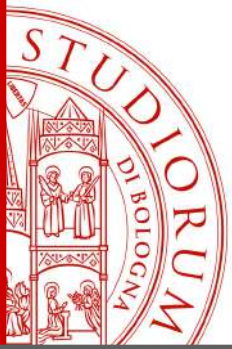
## ARGOS MANAGES A SUBSET OF FLIGHTS

ARGOS manages certain flights (for each flight, a plan is presented and executed). The ATCO monitors and can take flights away from ARGOS. The ATCO controls all non-ARGOS flights.

L8

## ARGOS MANAGES ALL FLIGHTS

ARGOS manages all flights (for each flight, a plan is presented and executed). The ATCO is alerted by ARGOS when monitoring is required: ARGOS still manages the situation but outside its normal comfort zone (i.e. conflict-free look-ahead time is reduced). The ATCO monitors as requested (i.e. stays in L8). The ATCO can take flights away from ARGOS (i.e. revert to L5).



- Supporting flight and landing of aircraft operated by a single pilot, in case of partial or total incapacitation of the pilot.
- **Three implementation options:**



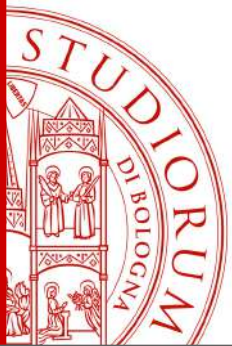
ATCO Focused:  
most of the single  
pilot tasks are  
assigned to the Air  
traffic controller



GSO Focused:  
most of the single  
pilot tasks are  
assigned to the  
Ground Station  
Operator

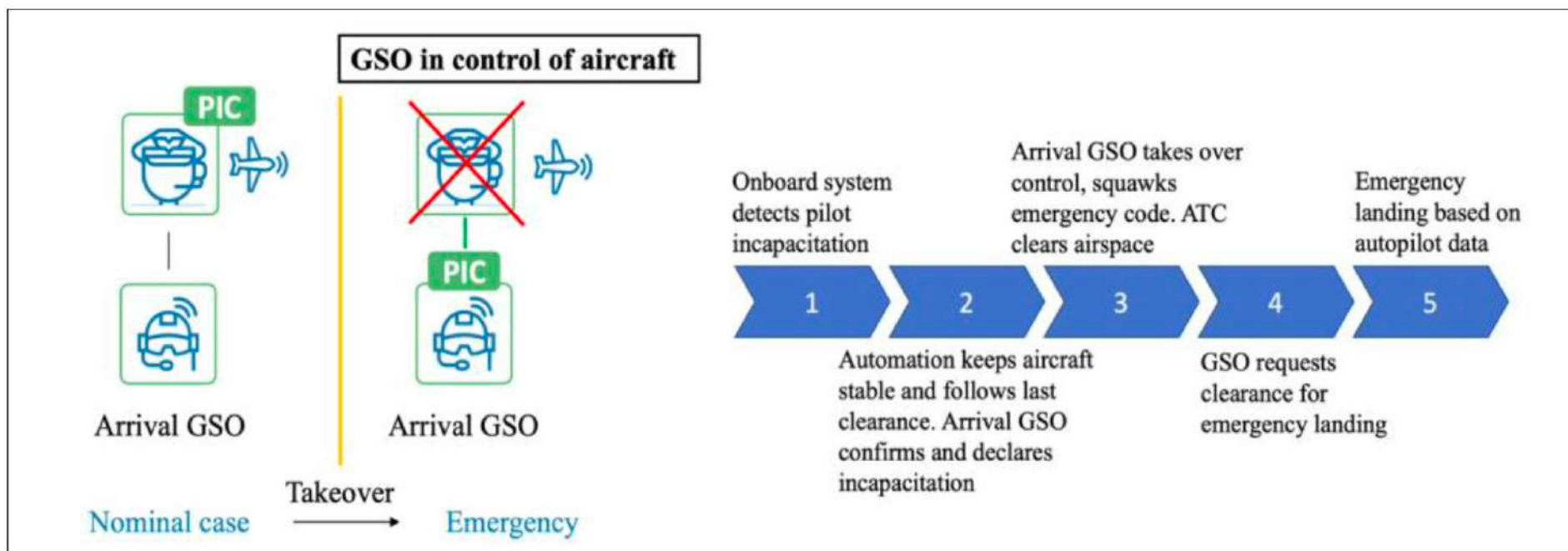


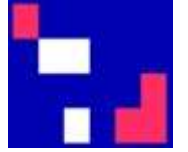
Automation  
Focused: most of  
the single pilot tasks  
are assigned to the  
cockpit automation



# SAFELAND

- Selected solution: GSO + Automation





# Deontology/Kantian ethics

Giovanni Sartor





# Deontology

- Consequentialists hold that choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about.
  - E.g. my act of lying is good or bad depending on the effects it brings in the world
- Deontologists hold that certain actions are good or bad regardless of their consequences
  - Lying is always bad, regardless of its effect.
- The right has priority over the good: what makes a choice right is its conformity with a moral norm which orders or permits it, rather than its good or bad effect.
  - E.g. we should not kill anybody, even in those cases in which killing somebody would provide more utility. Is this always the case
    - Consider the case of the British soldier who apparently met Hitler in the trenches of 1<sup>st</sup> world war
    - What would a rule utilitarian say in such a case?
- The 10 commandments?

# Some ideas for being impartial

## Ethics and impartiality

- Is ethics linked to ideas of fairness or impartiality?
- Is it unethical to have a preference for oneself (or one's friends)?

## What about the golden rule

- Treat others as you would like others to treat you
- Do *not* treat others in ways that you would *not* like to be treated
- What you wish upon others, you wish upon yourself

## Is the golden rule useful

- Always? Can you find counterexamples?
- Would you want an AI system that applies it (with regard to its owner)?

# Immanuel Kant

- One of the greatest philosophers of all times
- Lived in Prussia (1724-1804)
- Addressed
  - The theory of knowledge: Critique of pure reason
  - The theory of morality: Critique of practical reasons
  - The theory of aesthetics (art): Critique of judgment
  - Law, logic, astronomy, etc.



# Kant's ethic and the principle of universalizability

- “Act only according to that maxim by which you can at the same time will that it should become a universal law” (1785).
- What is a maxim: a subjective principle of action, it connects an action to the reasons for the action (an intention to perform an action for a certain reason)
  - I shall donate to charities to reduce hunger
  - I shall deceive my contractual partner, to increase my gains
  - I shall cheat on taxes, to keep my money
  - I shall tell the truth, to provide trust
- Are they universalizable? Would I want them to become universal laws, that are applied by everybody?

# An universalisation test

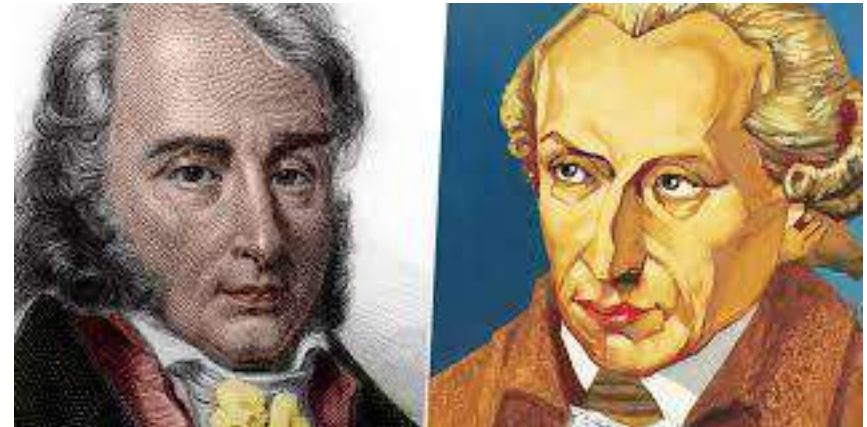
- Shafer Landau. The test of universalizability:
  - Formulate your maxim clearly state what you intend to do, and why you intend to do it.
  - Imagine a world in which everyone supports and acts on your maxim.
  - Then ask: Can the goal of my action be achieved in such a world?
- The process ensure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

Apply this principle to

- Cheating in an exam, in order go get a good mark
  - Giving money to a charity to relieve
- Would we want a robot following this maxim?

# Immanuel Kant vs Benjamin Constant

- Should one must (if asked) tell a known murderer the location of his prey.
  - It is ok to refuse to answer?
  - It is ok to tell a lie (e.g., if threatened by the murderer)?
- Is the maxim of telling lies universalizable?
- Is it defeasible?
- Its it Ok to have a robot that tells lies:
  - What about Asimov Liar
  - What about HAL in



# Hypothetical imperatives

- Hypothetical imperative: they require us to do what fits our goals
  - I would like to have more money
  - If cheat on taxes I will have more money
  - I shall cheat on taxes to have more money
  
- I would like to get a good mark
  - If I study I will get a good mark
  - I shall study
  
- Is this OK?
- The imperative is dependent on what I want (getting good marks, having more money)
  - I shall cheat on taxes, to having more money!

# The categorical imperative

- A moral imperative that applies to all rational beings, irrespective of their personal wants and desires,
- “Act only on that maxim through which you can at the same time will that it should become a universal law”
  - - make false premises when it suits you to do so?
  - - refuse help to do those who are in need when it suits you to do so?



# The good will

- The morality of an action only depends only to the extent that this action is motivate by our good will, i.e., by the necessity to comply with the categorical imperative
  - E.g., if I do well my job only in order to get a promotion, or be better paid I am not acting morally
  - I am acting morally if I do well my job because I think that this is my categorical duty, since I believe that everybody should act upon the maxim that they ought to do well their job to ensure societal progress
- The good will is the only thing that is good in itself
  - Do you agree?

# Another version of the categorical imperative: the principle of humanity

- So act that you treat humanity in your own person and in the person of everyone else always at the same time as an end and never merely as means
  - How is it linked to universalizability: As you consider your self as an end, you should consider the others in the same way (universalizability)?
- What does it mean treating somebody as an end (not as a mere means)
  - It cannot mean that we never use people for our purposes (e.g., when we ask for favours or pay for jobs)
  - It must mean that we should never treat people ONLY as means, without considering their values and purposes

# When does AI treat people only as means

- Autonomous weapons?
- Deceiving advertisements?
- Discriminatory appointments?
  
- When does AI fail to recognise humans as valuable entities, that should achieve their aims according to their choices?
  
- Can we treat AI systems only as means?

# Dignity

- For Kant rational beings, capable of morality (humans) have a special status “an intrinsic worth, i.e., **dignity**,” which makes them valuable' “above all price
  - Because of dignity they deserve respect
  - They cannot be treated as mere ends
- What does it mean that AI systems should respect human dignity, respect humans

# The foundations of dignity

- Why do humans deserve dignity. Because they have
  - Reason: they act on reasons and are aware of this
  - Autonomy: they can choose what to do, and in particular to follow the categorical imperative rather than their subjective preference
- 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
- What if AI system also had reason and autonomy
- Would they become citizens of the kingdom of ends

# Morality as an aspect of rationality

- For Kant if we follow rationality, we have to be moral.
  - Can there be a rational criminal?
  - It is rational to pursue my wellbeing at the expense of others?
  - Is it rational for a company to develop a system that is profitable, but that will cause more harm than good (e.g.,

# Rationality and consistency

- 1. If you are rational, then you are consistent.
- 2. If you are consistent, then you obey the principle of universalizability.
- 3. If you obey the principle of universalizability, then you act morally.
- 4. Therefore, if you are rational, then you act morally.
- 5. Therefore, if you act immorally, then you are irrational.

What kind of consistency is this?

- If I deserve something no less than others, and I want it for me, I should recognise it also to others!
- Is this consistent with rationality? Is it required by it? Can I be rational, and pursue my goal to the detriment of other

# Issues

- Does the principle of universalizability always provide acceptable outcomes
- Is it sufficient that the maxim of my action is such that I would like it to be universalised for this maxim to be good?
- Can you think of some examples when this is not the case?
  - Lying ? Robbing? Celibacy? Genocide?



# Alan Gewirth: principle of generic consistency

1. I do (or intend to do) X voluntarily for a purpose E that I have chosen.
2. E is good
3. There are generic needs of agency.
4. My having the generic needs is good *for* my achieving E *whatever E might be*  $\equiv$  My having the generic needs is categorically instrumentally good for me.<sup>13</sup>
5. I categorically instrumentally ought to pursue my having the generic needs.
6. Other agents categorically ought not to interfere with my having the generic needs *against my will*, and ought to aid me to secure the generic needs when I cannot do so by my own unaided efforts *if I so wish*,
7. I am an agent  $\rightarrow$  I have the generic rights.
8. All agents have the generic rights.

Other attempts exist to develop a Kantian ethics.

# 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

# David Ross (1877 1971): prima facie duties

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

# Defeasibility of duties

- Does it make sense to view duties as being defeasible?
- Can we apply defeasible reasoning to reason with duties?
- Should an AI system admit exceptions to duties, or should it always ask humans?

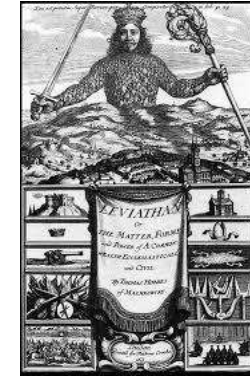
# Contractarianism

Giovanni Sartor

# 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)

# State of nature and social contract

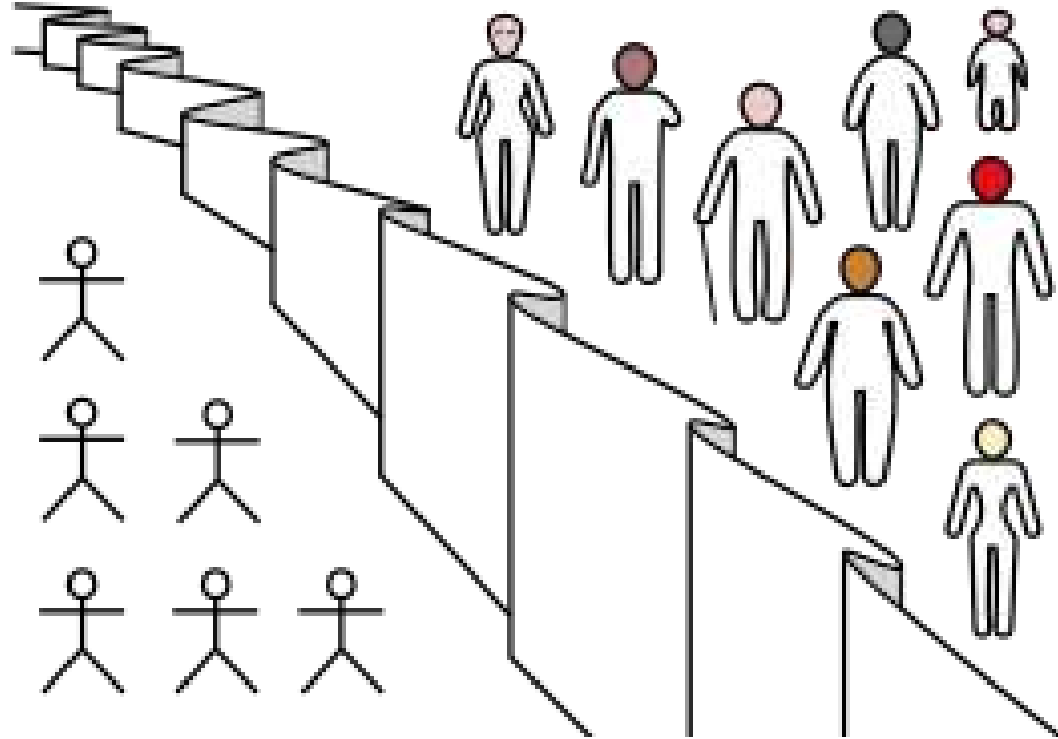


- How to get out of the state of nature?
- What agreements are OK?

		B	
		Cooperate	Defect
A	Cooperate	4, 4	-2, 6
	Defect	6, -2	0, 0

# John Rawls (1921-2002)

- A theory of justice
- How to ensure that the social contract is fair?
- People should choose under a **veil of ignorance**, without knowing their gender, social position, interests talents, wealth, race, etc.





# What principles would they go for?

- **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*). (JF, 42–43)

# AI in a just society (according to Rawls)

- Does the deployment of AI in today's society fit Rawls' requirements
- When may it conflict with the basic liberties?
- When with fair equality of opportunity?
- When with the difference principle?

# Juergen Habermas: Discourse Ethics

- A rule of action or choice is justified, and thus valid, only if all those affected by the rule or choice could accept it in a reasonable discourse.
- A norm is valid when the foreseeable consequences and side effects of its general observance for the interests and value orientations of each individual could be jointly accepted by all concerned without coercion
- The valid norms are those that would be the accepted outcome of an "ideal speech situation", in which all participants would be motivated solely by the desire to obtain a rational consensus and would evaluate each other's assertions solely on the basis of reason and evidence, being free of any physical and psychological coercion
- This approach assumes that people are able to engage in discourse and converge on the recognition of reasons for norms and choices

# Habermas and AI

- Would would we all agree if we engaged in an impartial discussion on how to use AI?
- Can we think of an AI system that engages in an impartial moral debate? What would it argue for?

# Virtue ethics

Giovanni Sartor

# Virtue ethics

- Ethics should not focus on norms nor on consequences
  - An act is morally right just because it is one that a virtuous person, acting in character, would do in that situation.
- Ethics is a complex matter
  - Since there are many virtues, the right act is that that would result from the mix of the relevant virtues: honesty; loyalty; courage; impartiality, wisdom, fidelity, generosity, compassion, etc.
- Ethics 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?
- How can we extract precise indications from an account of virtues and from virtuous examples? How much can we rely in tradition?
- What if virtues are in conflict?
- What are the paradigms of virtues to which we may refer to?

# AI and virtue ethics

- Should we, as developer of AI systems, be virtuous? What character traits should we cultivate in us?
- Should AI applications (AI agents be virtuous)?
- How can virtues be learned?
- If from example, can the training of an AI system lead to a virtuous behaviour of it?

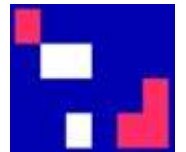


# Readings

- Shafer-Landau, R. (2018). The Fundamentals of Ethics. Oxford University Press.
- Singer, P. (2021). Ethics. In Encyclopedia Britannica: <https://www.britannica.com/topic/ethics-philosophy>

**MAI4CAREU**

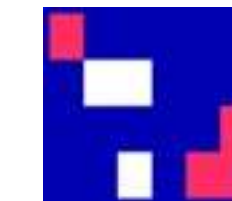
Master programmes in Artificial  
Intelligence 4 Careers in Europe



Co-financed by the European Union  
Connecting Europe Facility

This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423





# L'algoritmo umano

**Cosa significa rimettere al centro l'essere umano...  
al tempo dei robot.**

**Andrea Pezzi**



Co-financed by the European Union  
Connecting Europe Facility

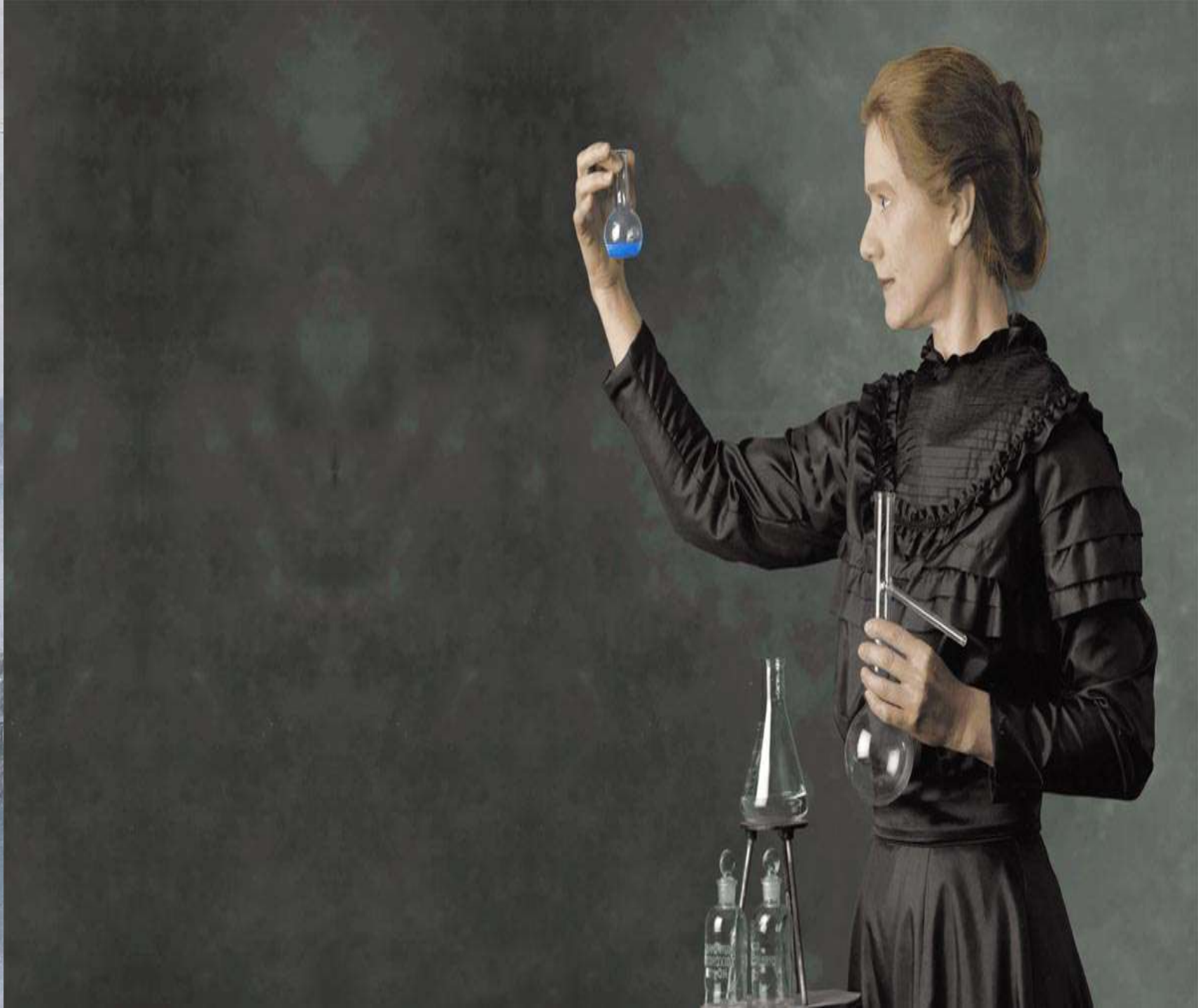
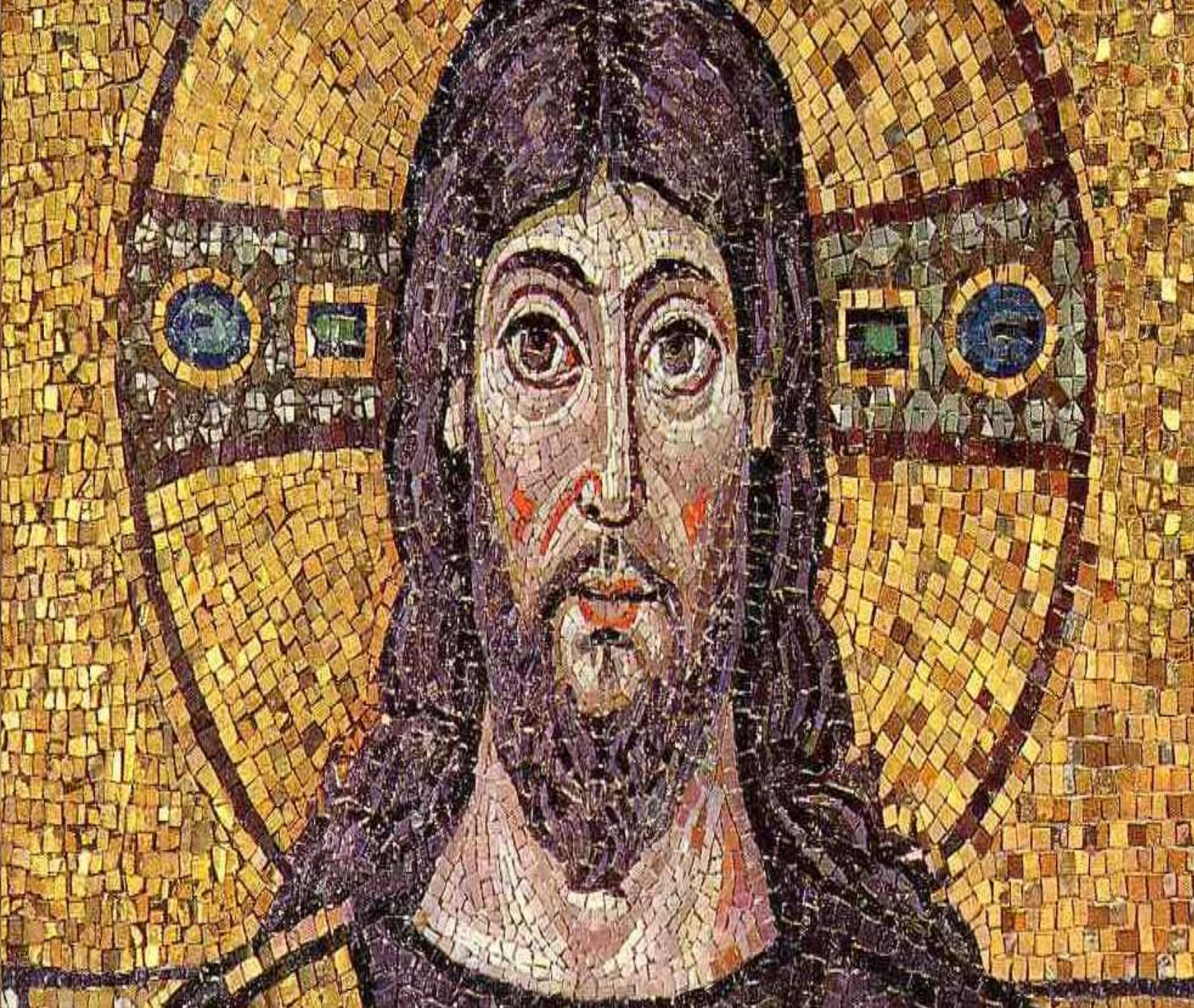
This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423



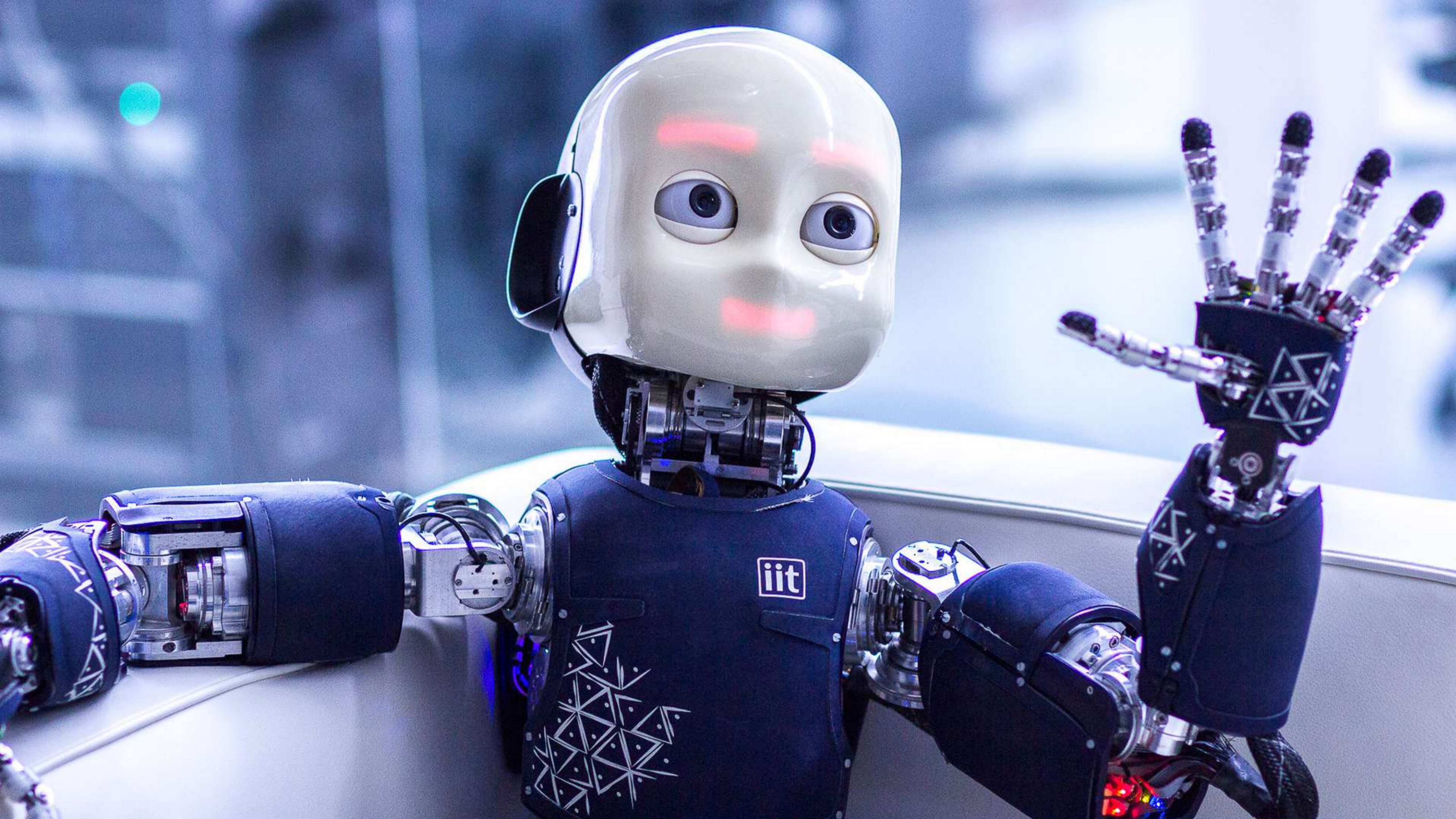








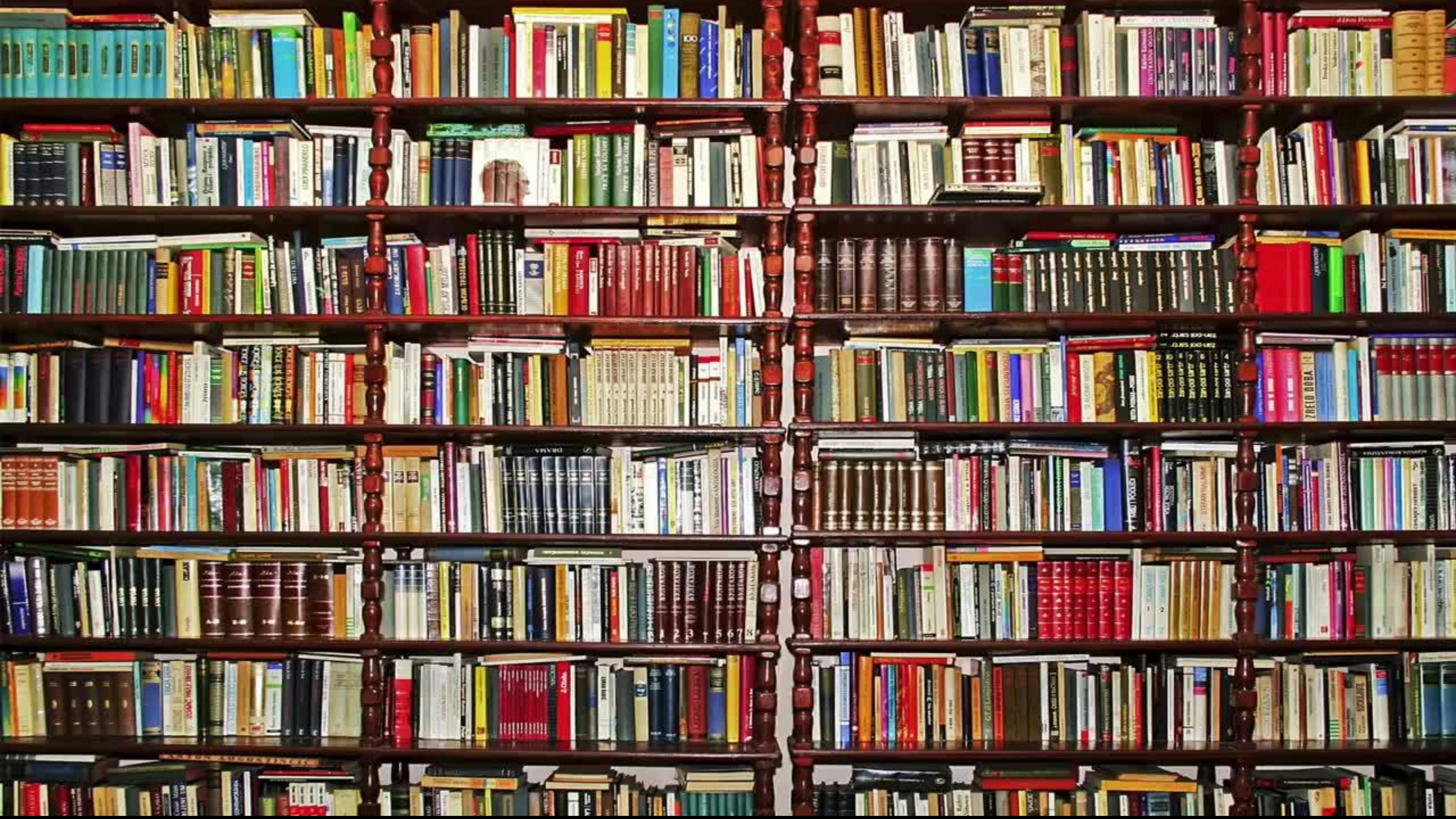






















# Get ready for Brexit

Prepare for Brexit at [gov.uk/brexit](http://gov.uk/brexit)

Für ein Deutschland, in dem wir gut und gerne leben.

## Ensemble, la France!

EMMANUEL MACRON

ELECTION PRÉSIDENTIELLE DU 7 MAI 2017

## CHOISIR LA FRANCE

MARINE PRÉSIDENTE

04 FEB. ORE 17.30 NUORO

PARTECIPARE SCEGLI. CAMBIA.

## PRIMA L'ITALIA!

IL BUONSENNO IN EUROPA

EUROPEE · DOMENICA 26 MAGGIO VOTA



legaonline.it · #primaitalia

BRASIL ACIMA DE TUDO, DEUS ACIMA DE TODOS!

VOTE 17

PRESIDENTE BOLSONARO  
vice General MOURÃO

COLIGAÇÃO BRASIL ACIMA DE TUDO, DEUS ACIMA DE TODOS. PRESIDENTE BOLSONARO, VICE PRESIDENTE MOURÃO

## TRUMP 2016

MAKE ★ AMERICA ★ GREAT ★ AGAIN

## Immreise

Jetzt **NPD**  
Die Nationalen

npd.de 030 - 650 110  
NPD, Postfach 84 01 57, 12531 Berlin



Palermo. Un 24 enne cittadino MALIANO ha aggredito e poliziotto e un funzionario.

Il migrante accusato di resistenza lesioni minacce violenze pubblico ufficiale è stato arrestato. Subito dopo l'apertura principale il giovane scavalcava le transenne per avventarsi contro il personale addetto alle pubbliche relazioni aggredendo le spalle afferrandolo al collo e colpendolo con pugni alla testa.

Nel frattempo altro personale dell'ufficio immigrazione in cerca di bloccare lo straniero, senza riuscirci.

L'aggressore infatti resisteva in maniera violenta all'azione cercando di divincolarsi dalla loro presa sferrando pugni.

Un agente cadeva per terra e lo straniero lo colpiva al viso che riportavano vistose ferite alla testa. Grazie all'intervento dei colleghi, è stato evitato il peggio e bloccato l'extracomunitario.

Per bloccarlo è stato utilizzato il taser e anche lo spray urticante. I due agenti così venivano trasportati con ambulanze al pronto soccorso.

Lo straniero sedato da personale medico sopraggiunto è stato trasportato da personale delle volanti presso le camere di detenzione.

"Le continue aggressioni che stiamo registrando nei confronti del personale in divisa - dice Giovanni Assenzio segretario generale provinciale Usip Palermo - non devono e non possono più essere tollerate, necessità che la politica dia un chiaro segnale di fermezza dotando gli operatori della sicurezza di strumenti e regole chiare. Non possiamo ancora assistere ad episodi di violenza nei confronti di questi servitori dello Stato".



BLOGSICILIA.IT  
**Palermo, migrante aggredito e mediatore culturale: portati in ospedale**

26 agosto Un 25enne nigeriano arrestato dai militari per tentata estorsione e violenza. Voleva 500 euro per lasciare il centro di accoglienza che lo aveva ospitato.

ULTIRRENO.GELOCAL.IT  
**Migrante furioso chiede soldi e aggredisce un carabiniere**  
Un 25enne nigeriano arrestato dai militari per tentata estorsione e violenza.

Un'ivoriano denuncia un'aggressione a suo carico, da parte di razzisti.. Ma dopo accurate indagini la polizia scopre che è una balla! Insomma... Si è inventato tutto! 😂😂😂  
È stato giustamente accusato di simulazione di reato! Ben gli sta! Magari gli entra in testa un po' di rispetto, cosa di cui dubito fortemente.

**Il Primato Nazionale**  
22 ottobre 2018

#Cronaca: l'uomo è stato denunciato con l'accusa di simulazione di reato

#Genova #immigrato #ivoriano #migrante #aggressione #razzismo #IlPrimatoN

ILPRIMATONAZIONALE.IT  
**Immigrato ivoiriano denuncia: "Aggredito dai razzisti". Ma si è inventato tutto**

PERIAPOST.IT  
**PERIAPOST.IT  
MIGRANTE AGGREDISCE CONTROLLORE SU BUS T. IL GIUDICE CONVALIDA L'ARRESTO E POI LO LIBERA/**

PERIAPOST.IT  
**PERIAPOST.IT  
MIGRANTE AGGREDISCE CONTROLLORE SU BUS T. IL GIUDICE CONVALIDA L'ARRESTO E POI LO LIBERA/**

Il taser uno strumento necessario!!!

BLOGSICILIA.IT  
**Palermo, migrante aggredisce con calci e pugni poliziotto e mediatore culturale: portati in ospedale | BlogSicilia -**

Ho i brividi se penso che chiunque, italiano, naturalizzato, migrante o richiedente asilo possa essere aggredito, nell'indifferenza generale, solo per il colore della pelle. Si chiama razzismo, un vecchio male la cui causa risiede esclusivamente nell'ignoranza e nel non essere mai usciti di casa. I governi, in queste situazioni, dovrebbero educare "il popolo" al rispetto e alla tolleranza. In queste emergenze invece, al momento, abbiamo un ministro della giustizia che incita all'odio razziale, un ministro della giustizia che propone norme ancora più severe contro i razzisti e un ministro della giustizia che parla solo se autorizzato. E questo è il nostro governo.

NAPOLI.FANPAGE.IT  
**Un abbraccio e il perdono: il migrante aggredito a Napoli incontra i baby aggressori**

ITALIA  
OSAKUE

Ennesima aggressione a migranti sfruttati Non è il mio paese!!!

Italia, 2018  
È questo in cui vorremmo vivere?

ANSA.IT  
**Migrante aggredito, pista razzista - Sardegna**  
Si segue la pista razzista per l'aggressione, avvenuta ieri sera a Fertilia,

Quando due tredicenni sparano ad un migrante e per tentare di difendere il loro gesto sminuendolo parlando di "goliardata" a perdere siamo tutti noi. Davanti a noi abbiamo un compito fondamentale, importantissimo ed urgente: spegnere il clima d'odio razziale che sta divampando nel paese, recuperare umanità.

La parrocchia di Don Massimo Biancalani a Vicofaro, in provincia di Pistoia, da anni realizza ottimi progetti di integrazione: sparare ad un migrante accolto in una parrocchia è un gesto che ha significati ben precisi, perché significa colpire uno dei tanti esempi concreti di integrazione riuscita in Italia.

Solidarietà a Don Biancalani e a chi è impegnato tutti i giorni a promuovere una convivenza pacifica.

FIRENZE.REPUBBLICA.IT  
**Pistoia, sono stati due tredicenni a sparare al migrante**  
L'aggressione a un ragazzo del Gambia di 24 anni ospite della parrocchia

Alghero non vuole la feccia razzista, non la tollera e non la rappresenta: Preso a calci e pugni un ragazzo senegalese al grido di "Ti investo, tanto ti ammazzo, sporco negro" "Tornatevene al vostro Paese"

E che tristezza ho provato oggi quando anche io, "straniero" in Gallura, sono stato oggetto di una frase di questo genere.

ANSA.IT  
**Migrante aggredito, pista razzista - Sardegna**  
Si segue la pista razzista per l'aggressione, avvenuta ieri sera a Fertilia,

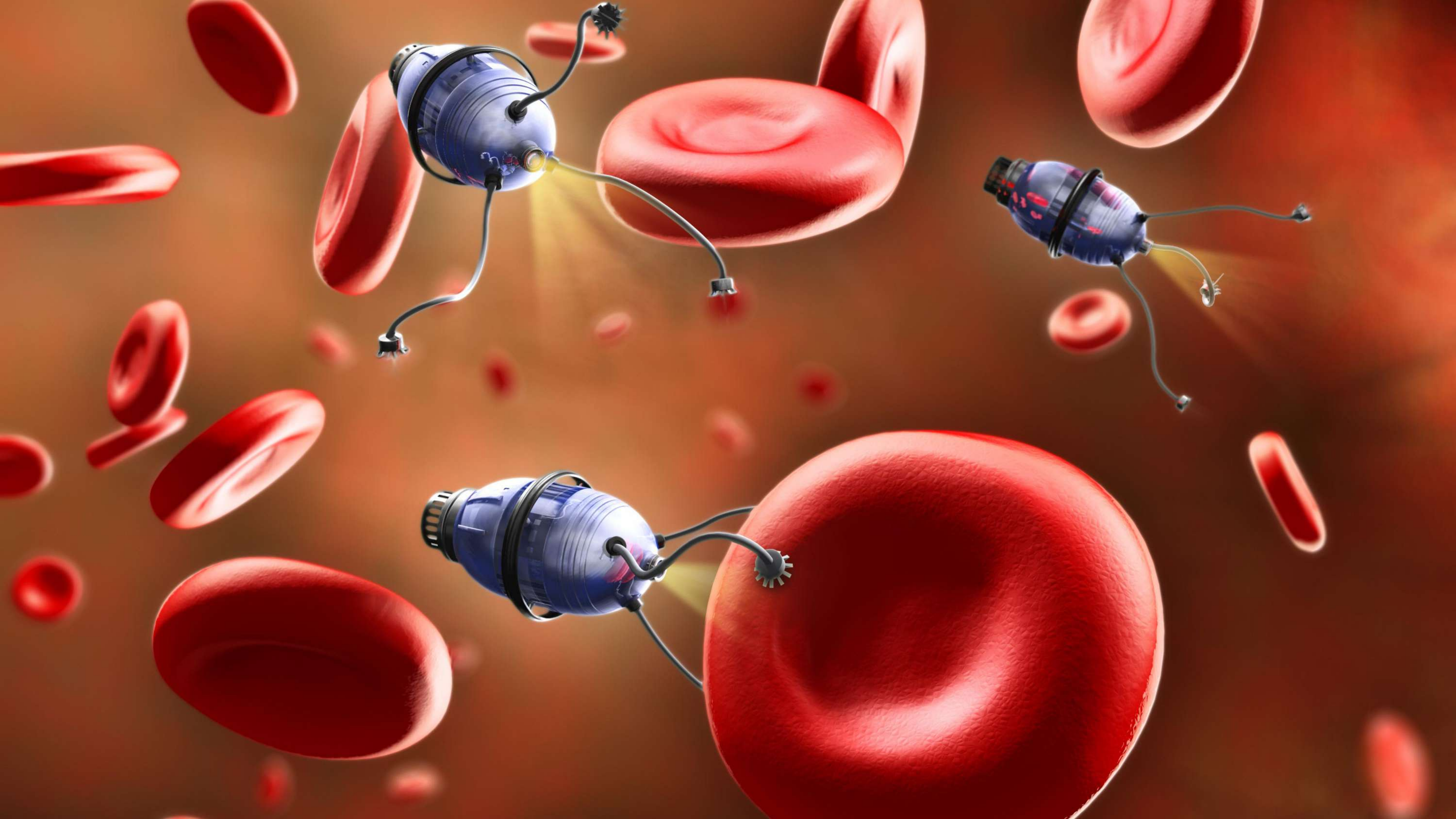
Quando due tredicenni sparano ad un migrante e per tentare di difendere il loro gesto sminuendolo parlando di "goliardata" a perdere siamo tutti noi. Davanti a noi abbiamo un compito fondamentale, importantissimo ed urgente: spegnere il clima d'odio razziale che sta divampando nel paese, recuperare umanità.

La parrocchia di Don Massimo Biancalani a Vicofaro, in provincia di Pistoia, da anni realizza ottimi progetti di integrazione: sparare ad un migrante accolto in una parrocchia è un gesto che ha significati ben precisi, perché significa colpire uno dei tanti esempi concreti di integrazione riuscita in Italia.

Solidarietà a Don Biancalani e a chi è impegnato tutti i giorni a promuovere una convivenza pacifica.

FIRENZE.REPUBBLICA.IT  
**Pistoia, sono stati due tredicenni a sparare al migrante**  
L'aggressione a un ragazzo del Gambia di 24 anni ospite della parrocchia





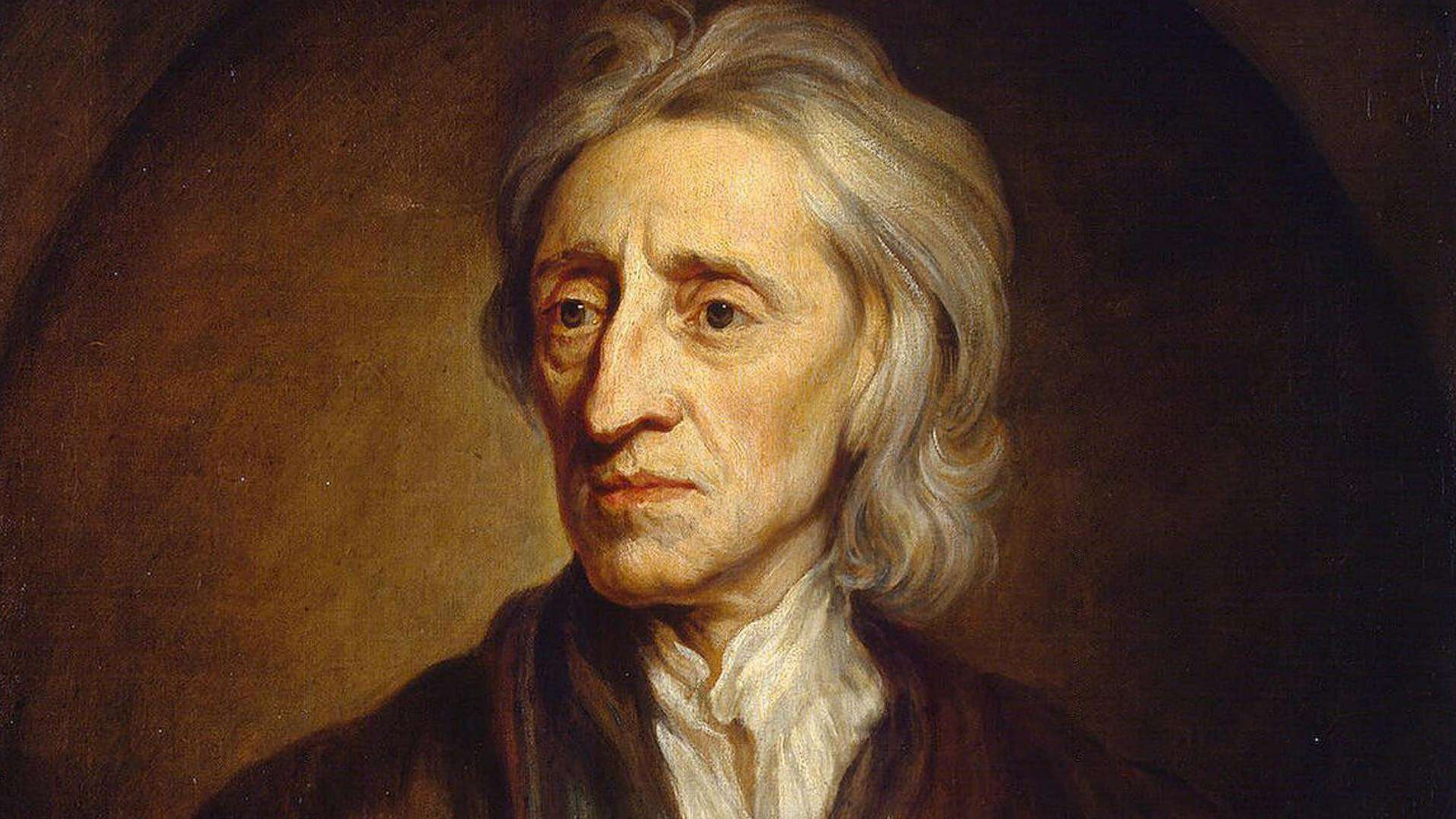




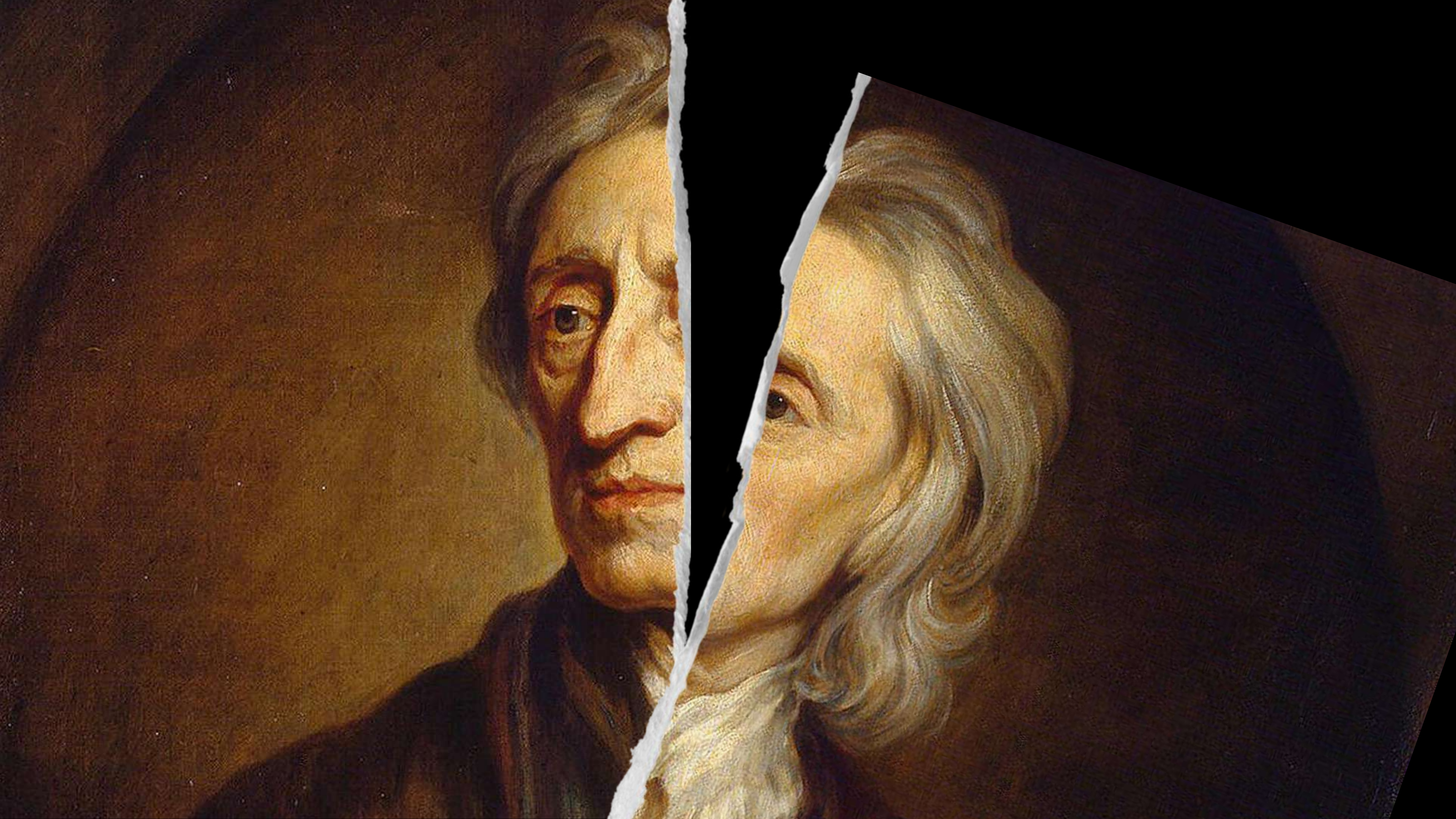




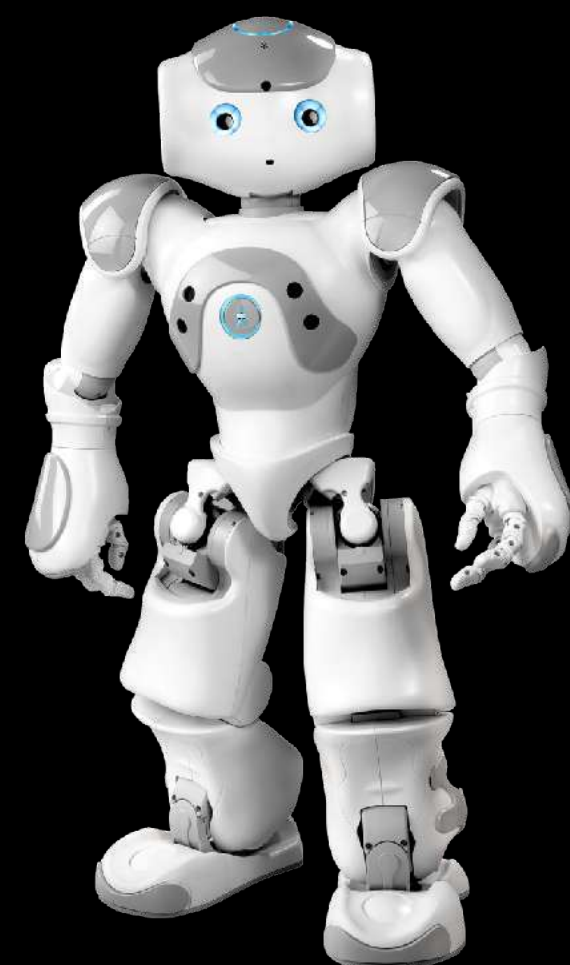












# PROGRAMMA

Dal latino *PRO GRAPHO*:  
scrivo prima



# PROGETTO

Dal latino *PRO JECTUS*:  
azione di gettare avanti

**ACCESO**

**SPENTO**



0

1



# PITAGORA DI SAMO

(570 a.C. - 490 a.C.)



# PITAGORA DI SAMO

(570 a.C. - 490 a.C.)

1



Punto  
(Prima dimensione)



# PITAGORA DI SAMO

(570 a.C. - 490 a.C.)

2



Linea  
(Seconda dimensione)

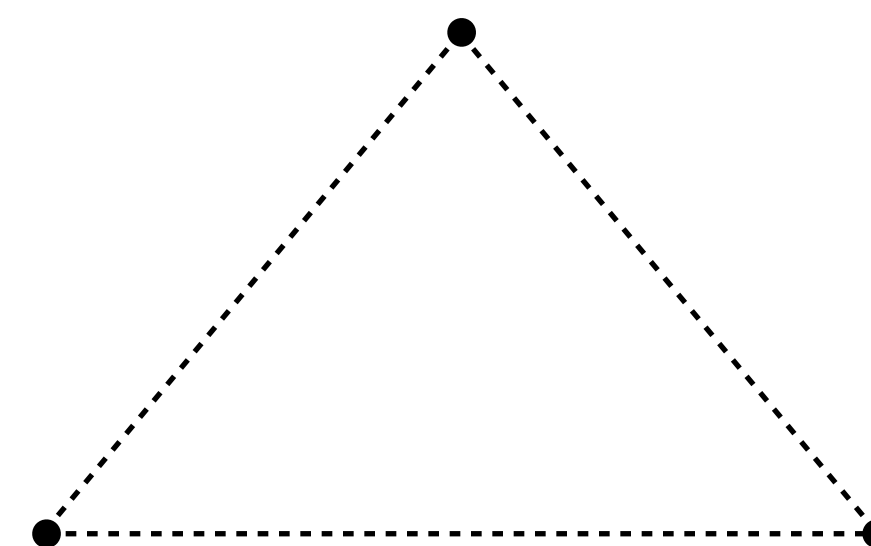




# PITAGORA DI SAMO

(570 a.C. - 490 a.C.)

3



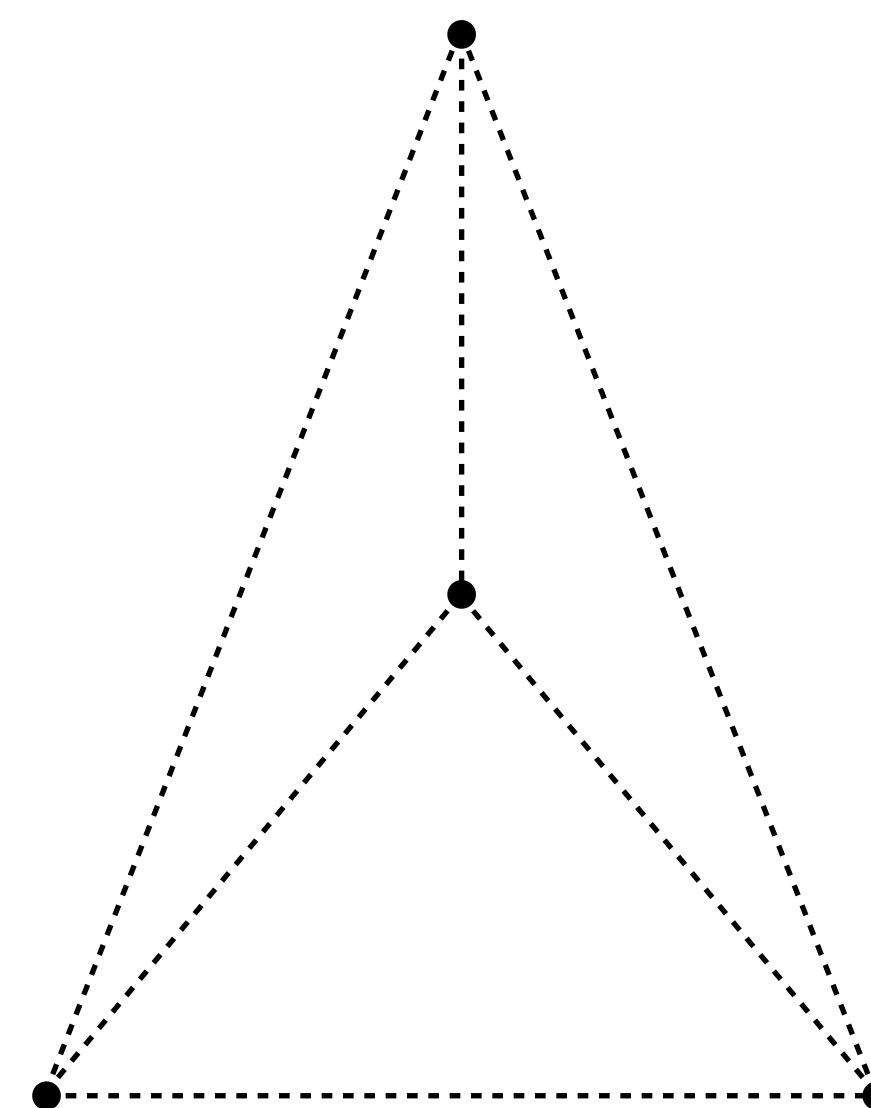
Superficie  
(Seconda dimensione)



# PITAGORA DI SAMO

(570 a.C. - 490 a.C.)

4



Solido  
(Terza dimensione)



**PITAGORA DI SAMO**

(570 a.C. - 490 a.C.)

**1**

**PARIMPARI**





# IPPASO DI METAPONTO

(450 a.C. - 485 a.C.)





**PITAGORA DI SAMO**

(570 a.C. - 490 a.C.)

COSTANTE DI PITAGORA

$$\sqrt{2} = 1,414213562373095\dots$$

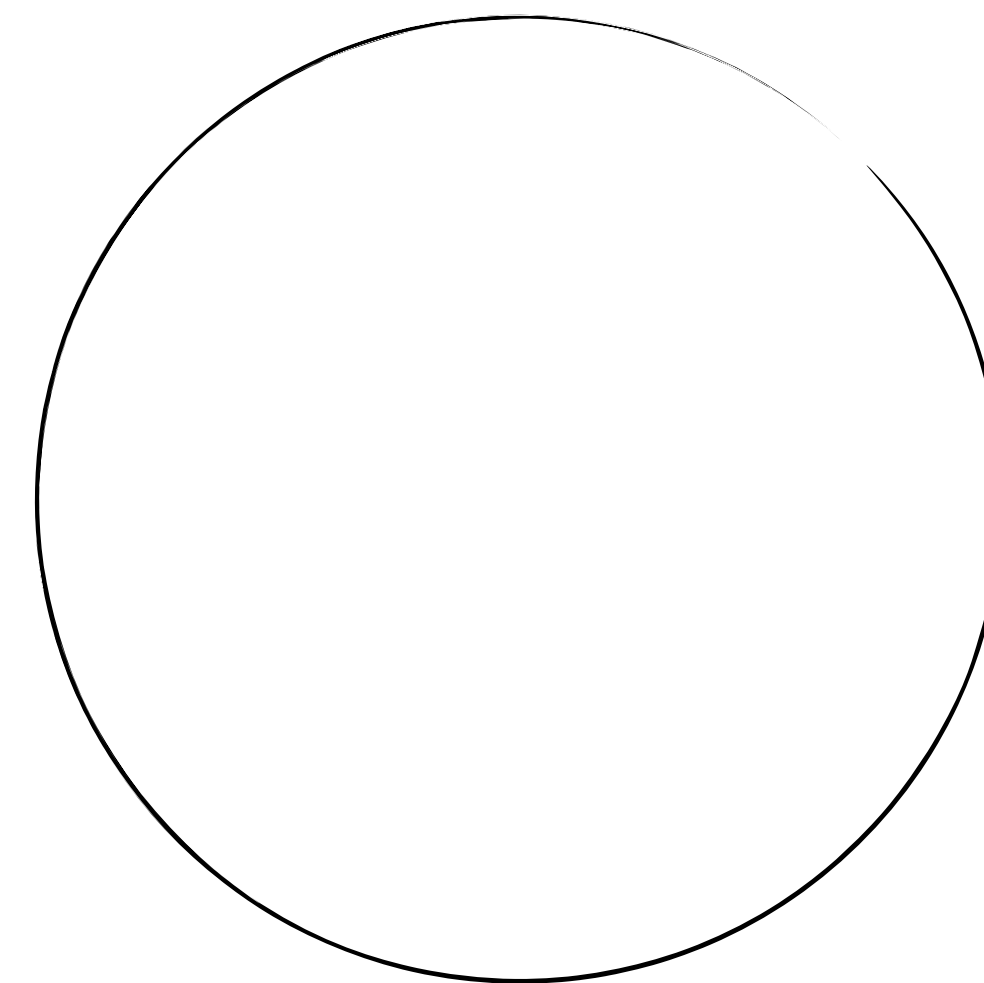




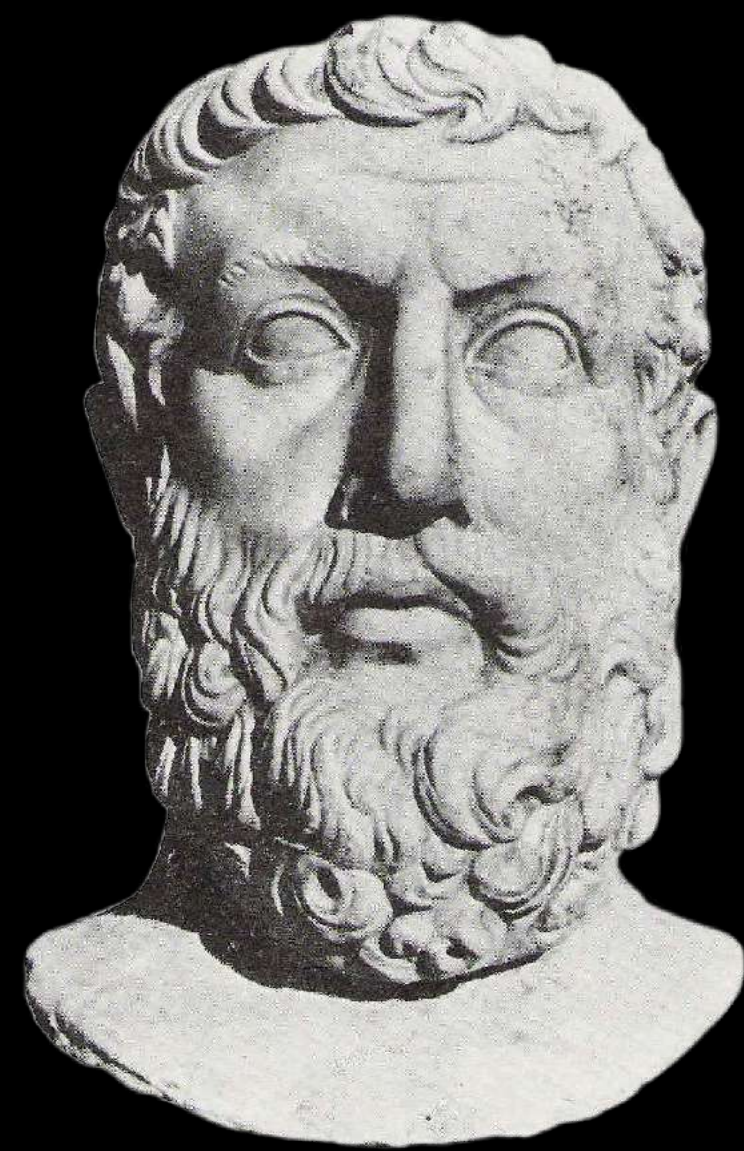
$$1 + 1 =$$

1

**DA DOVE  
VIENE  
L'ASSOLUTO  
DENTRO  
DI NOI?**







# PARMENIDE DI ELEA

(515 a.C. - 450 a.C.)



**PARMENIDE DI ELEA**

(515 a.C. - 450 a.C.)

**“  
L’essere è,  
il non essere non è  
”**





**PARMENIDE DI ELEA**

(515 a.C. - 450 a.C.)

**Aletheia**

**Doxa**

0

**Doxa**

1

**Aletheia**







**FORMA**

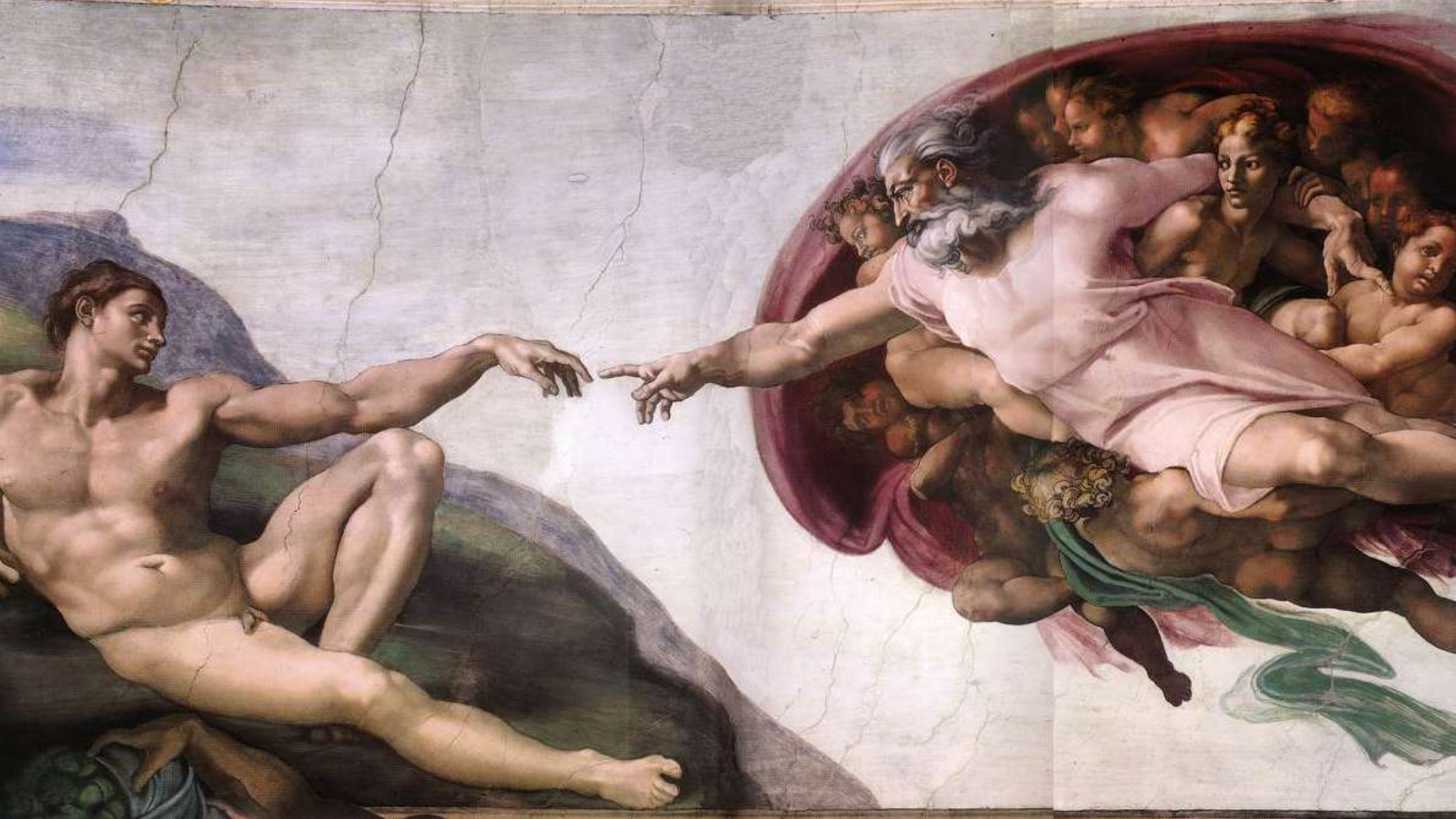
**SOSTANZA**

MESSAGGIO

**IMMAGINE**

**ENERGIA**



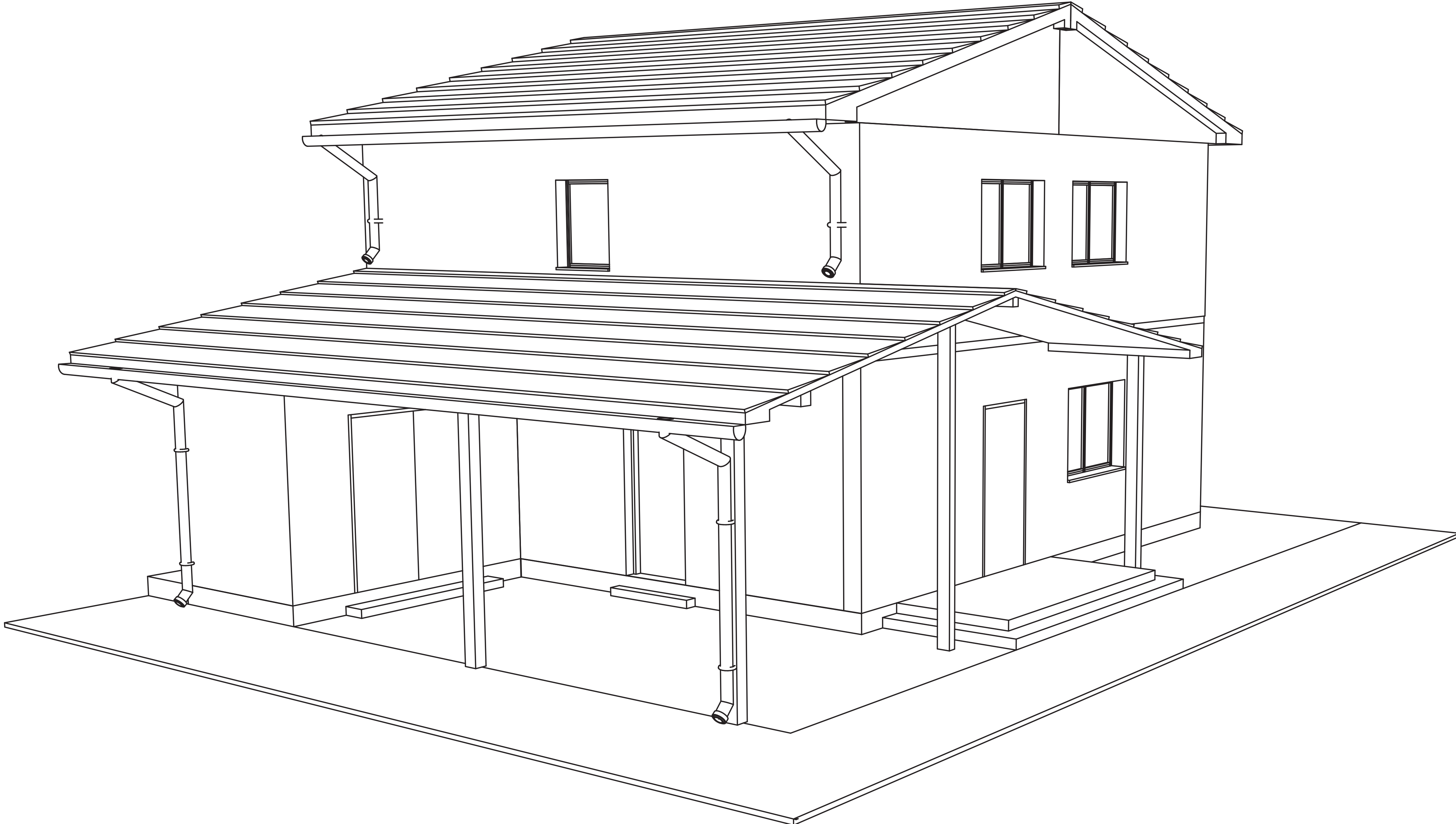


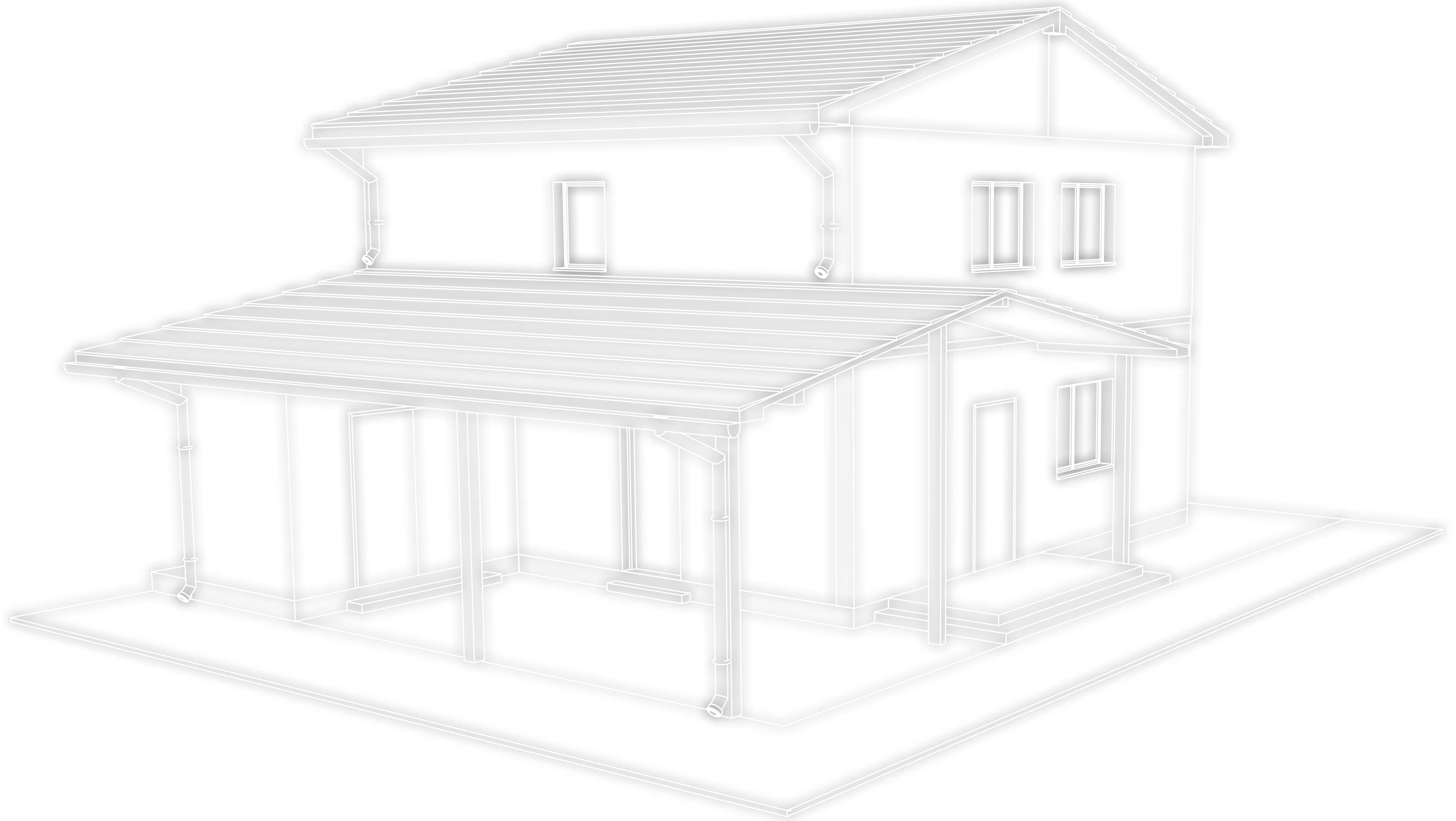


NON UN CODICE DI PROGRAMMA,  
MA UN PROGETTO VIVENTE,  
FATTO DI IMMAGINI INVISIBILI.











È DALL'INVISIBILE  
CHE NASCE  
IL VISIBILE.





L'IMMAGINE  
È ORDINE  
DELLA FORZA

L'IMMAGINE  
È ALFABETO  
DELL'ENERGIA



**ERACLITO**

(535 a.C. - 475 a.C.)





**ERACLITO**

(535 a.C. - 475 a.C.)

“*Panta rei,*”



**ERACLITO**

(535 a.C. - 475 a.C.)

“*Logos*”



**ERACLITO**

(535 a.C. - 475 a.C.)

“  
*Il Logos significa  
parola, verbo ma  
anche legge intesa  
come legge interiore,  
il logos interiore è  
ciò che va ascoltato  
per svegliarsi alla  
verità*  
”





**ERACLITO**

(535 a.C. - 475 a.C.)

**“*Il Logos è una sorta  
di ordine universale  
del divenire*”**



**ERACLITO**

(535 a.C. - 475 a.C.)

**“*Ad ogni uomo è  
concesso di  
conoscere se stesso  
ed essere saggio*”**

ONTOLOGIA

*Tendo a cercare  
la purezza*

*Tendo a cercare  
l'uguaglianza con  
l'altro*

*Tendo a conoscere  
me stesso*

VERITÀ

ETICA

*E' giusto  
voler essere  
puri*

*E' giusto voler  
essere uguali*

*E' giusto voler  
conoscere se  
stessi*

BENE

ESTETICA

*La vita  
è purezza*

*La vita  
è uguaglianza*

*La vita  
è conoscenza  
di sé*

BELLEZZA





**ERACLITO**

(535 a.C. - 475 a.C.)

“*Chi eleva un discorso  
individuale a verità  
sociale è sordo al  
logos universale*”

**ESTETICA**

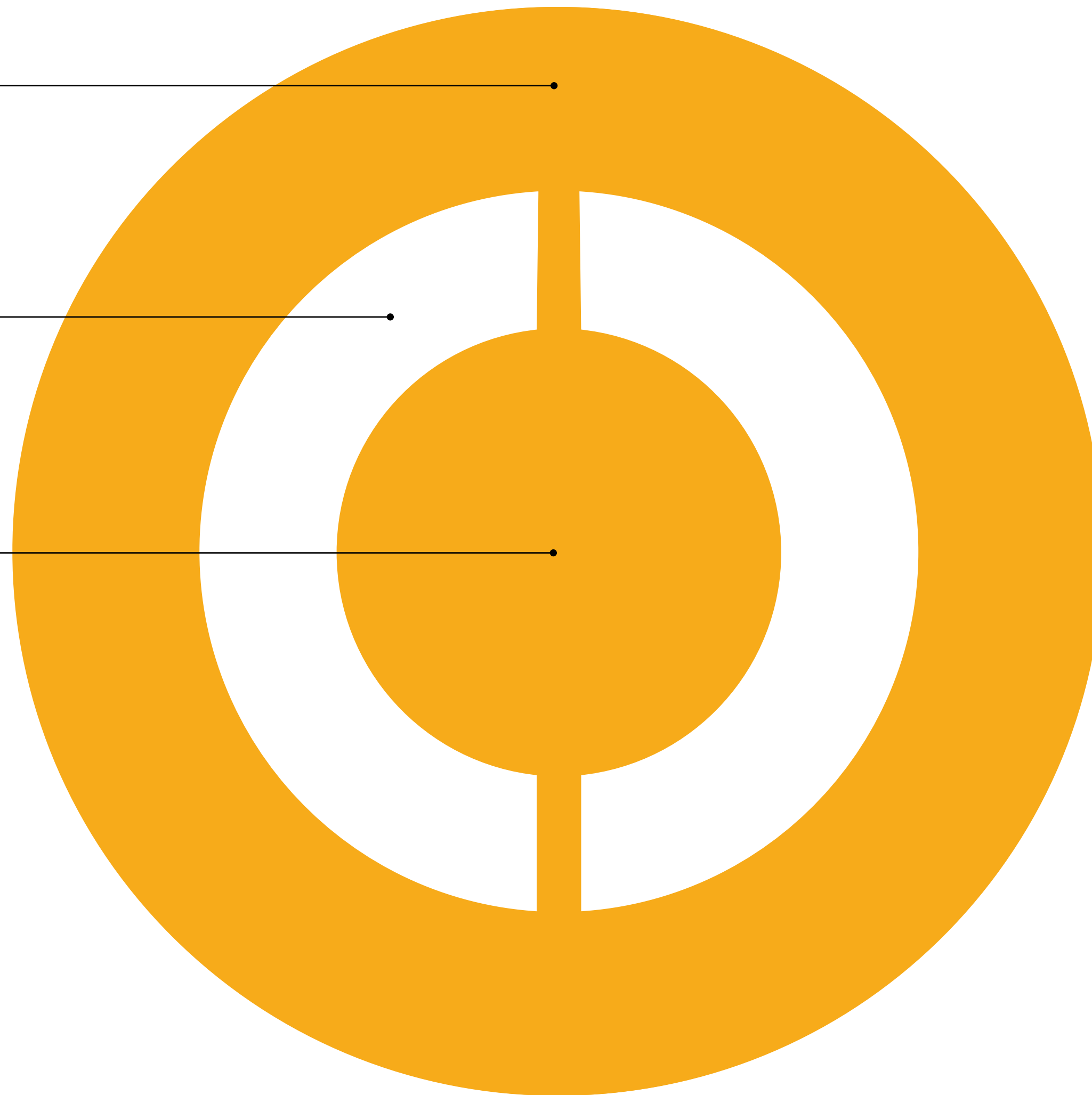
**ASSOLUTO**

**ETICA**

RELATIVO

**ONTOLOGIA**

**ASSOLUTO**





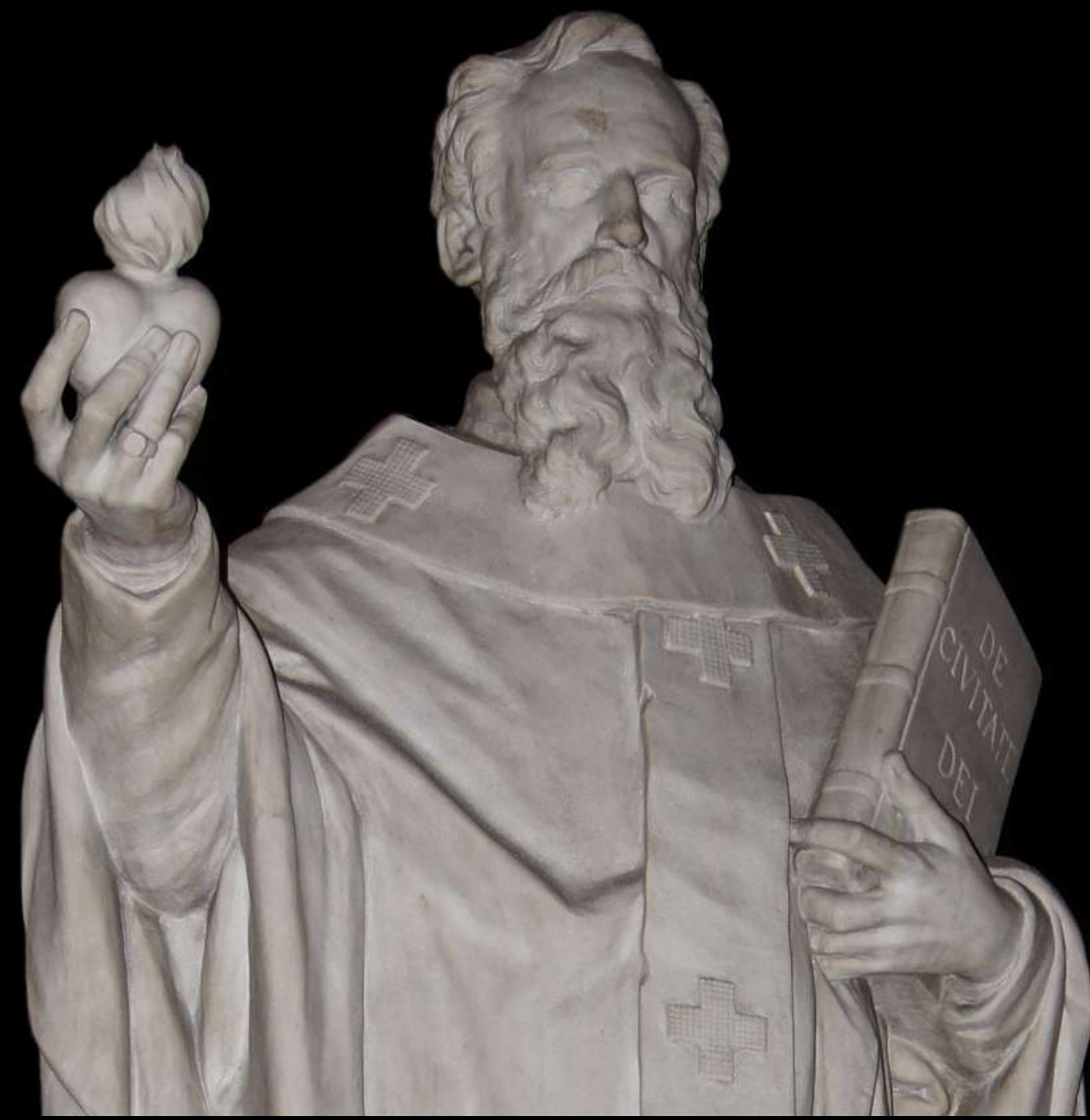
ARBEIT MACHT FREIHEIT











**SANT'AGOSTINO**

(354 d.C. - 430 d.C.)

**QUID EST VERITAS?**







SONO

so



FACCIO

SONO

so



FACCIO

SONO

so



FACCIO

SONNO

so

ACCIDENT



ON

S

O

NOCI

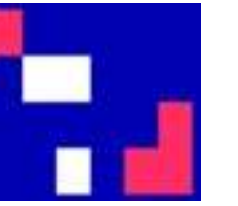










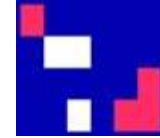


# IO SONO

GLI ALTRI PER INCONTRARE ME STESSO







# A GENETIC APPROACH TO THE ETHICAL KNOB

Giovanni IACCA (University of Trento)  
Francesca LAGIOIA (EUI/CIRSFID)  
Andrea LOREGGIA (EUI)  
Giovanni SARTOR (EUI/CIRSFID)





# AUTONOMOUS VEHICLES



# AUTONOMOUS VEHICLES

- **Autonomous Driving is classified according to the amount of human driver intervention:**
  - **From Level 0 (no automation) to Level 5 (full automation)**



# AUTOMATION LEVELS OF AUTONOMOUS CARS

## LEVEL 0



There are no autonomous features.

## LEVEL 1



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

## LEVEL 2



These cars would have at least two automated functions.

## LEVEL 3



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

## LEVEL 4



These cars are officially driverless in certain environments.

## LEVEL 5



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





# **AUTONOMOUS VEHICLES**

The amount of data to process increase with the level of automation

- **4.4 GB/s Data Logging for full Autonomous Driving**

## CAR AUTOMATION SENSORS & DATA VOLUMES

Sensor type	Quantity	Data generated
Radar	4–6	0.1–15 Mbit/s
LIDAR	1–5	20–100 Mbit/s
Camera	6–12	500–3,500 Mbit/s
Ultrasonic	8–16	<0.01 Mbit/s
Vehicle motion, GNSS, IMU	-	<0.1 Mbit/s

### TOTAL ESTIMATED BANDWIDTH

**3 Gbit/s (~1.4TB/h) to 40 Gbit/s (~19 TB/h)**



**AUTONOMOUS  
VEHICLES CAN  
POTENTIALLY FAIL**



# OUTLINE

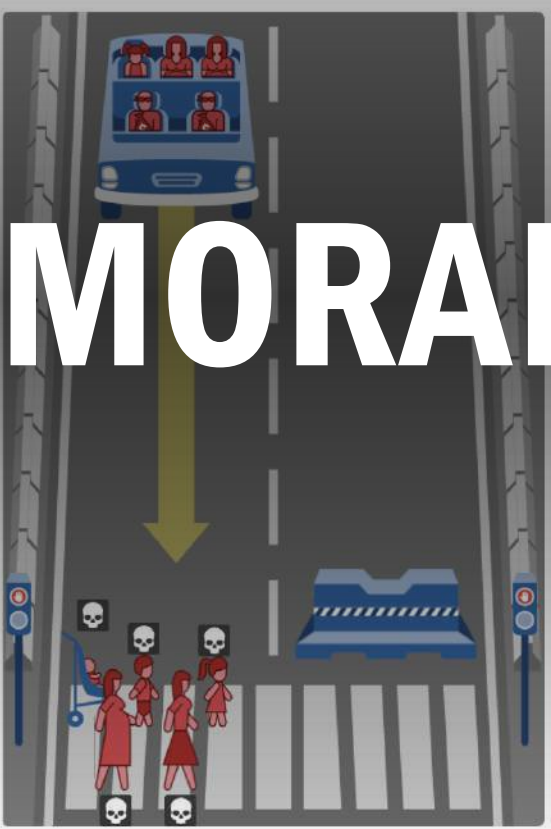
- **Introduction**
- **Ethical knob, individual preferences and social values**
- **Genetic Algorithms**
- **Neural Networks**
- **Genetic Approach to the Ethical Knob**
- **Conclusion/Discussion**



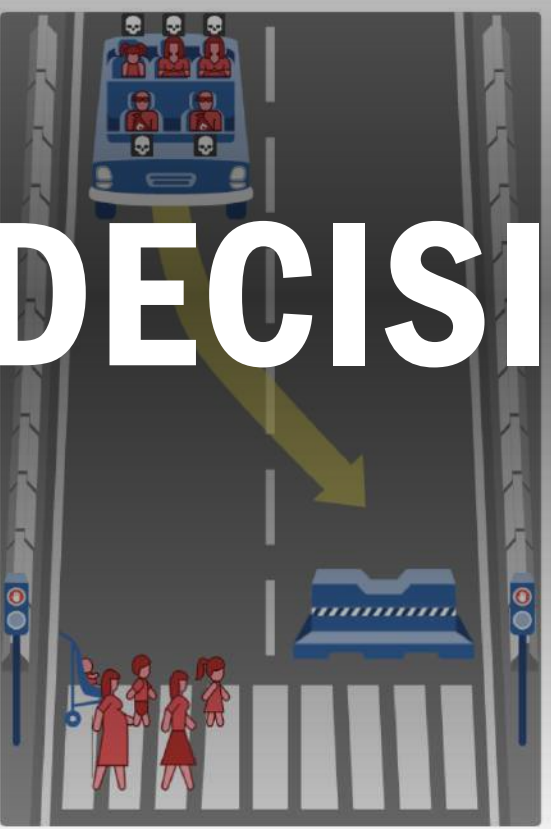
# Help

Share Link 0 Likes Random

# THE MORAL DECISIONS



Show Description



Show Description

# THE ORIG. PROPOSAL

- The knob expresses directly the ethical attitude of the AV passengers
- The value passengers attribute to their life relative to the value of the lives of third parties

“ETHICAL KNOB” SETTINGS

IMPARTIAL



STIC MODE: PREFERENCE TO PROTECT THE LIVES OF

SOURCE: CIRSFID, UNIVER

# THE NEW PROPOSAL

- The position of the knob no longer indicates the passengers' moral attitude
- It indicates the AV's assessment of the relative importance of the lives of passenger(s) and third parties

“ETHICAL KNOB” SETTINGS

IMPARTIAL



STIC MODE: PREFERENCE TO PROTECT THE LIVES OF

SOURCE: CIRSFID, UNIVER

# HOW TO DO THAT?

- **Combination of AI techniques:**
  - Neural networks to compute the right action to take based on the given scenario
  - Genetic Algorithm to find an (almost) optimal configuration of neural networks



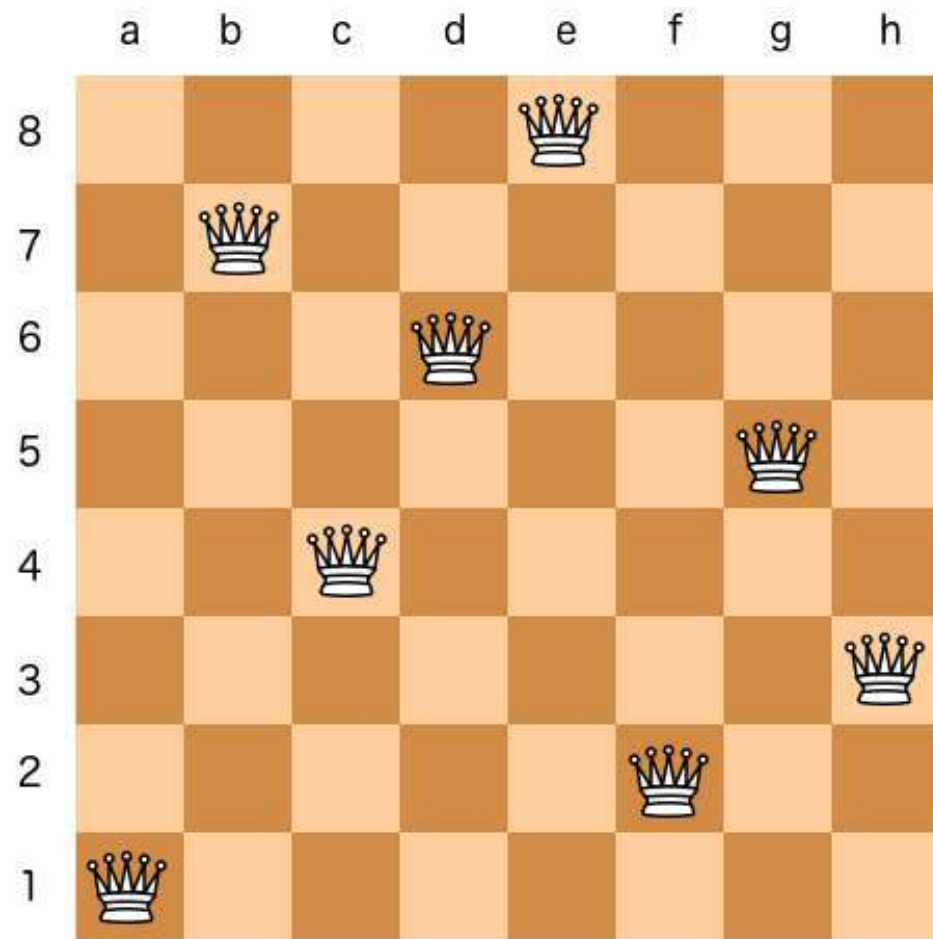
# GENETIC ALGORITHMS

- **Inspired by Charles Darwin's theory of natural evolution:**
  - the fittest individuals are selected for reproduction in order to produce offspring of the next generation
- **Heuristic Search in the solution space**
- **Mostly used in optimization tasks**

# SIMPLE EXAMPLE

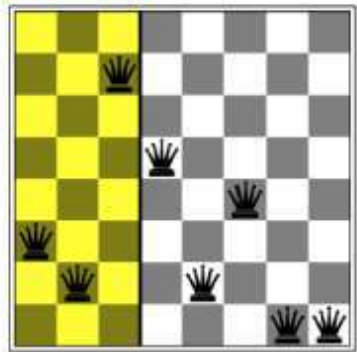
N-queens puzzle: place  $n$  chess queens on an  $n \times n$  chessboard so that no two queens threaten each other

# SIMPLE EXAMPLE

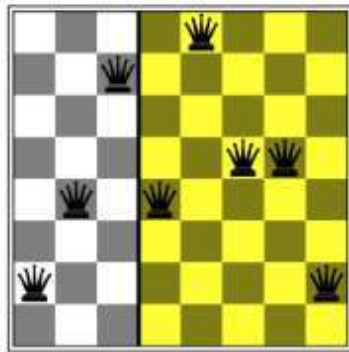


# HOW IT WORKS

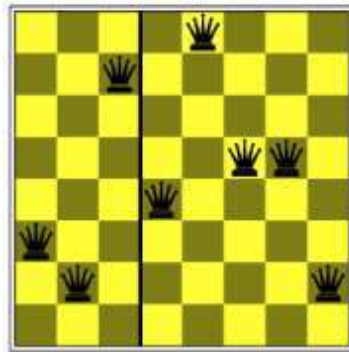
## Crossover



+



=



Taken from the edX course ColumbiaX: CSMM.101x Artificial Intelligence (AI)

- Individuals corresponds to solutions of the problem
- Initially, solutions are generated at random
- Each individual is evaluated
- The best are selected and used to combined to produce the new generation

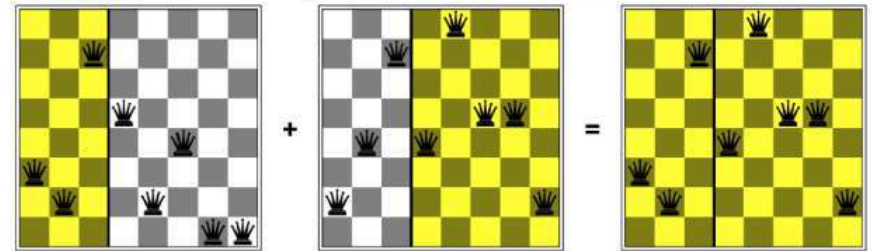


# HOW IT WORKS

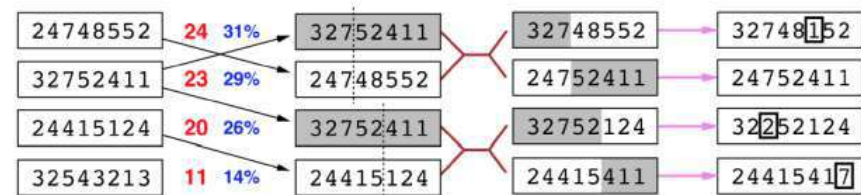
## Genetic algorithms

Crossover

Taken from the edX course ColumbiaX: CSMM.101x Artificial Intelligence (AI)

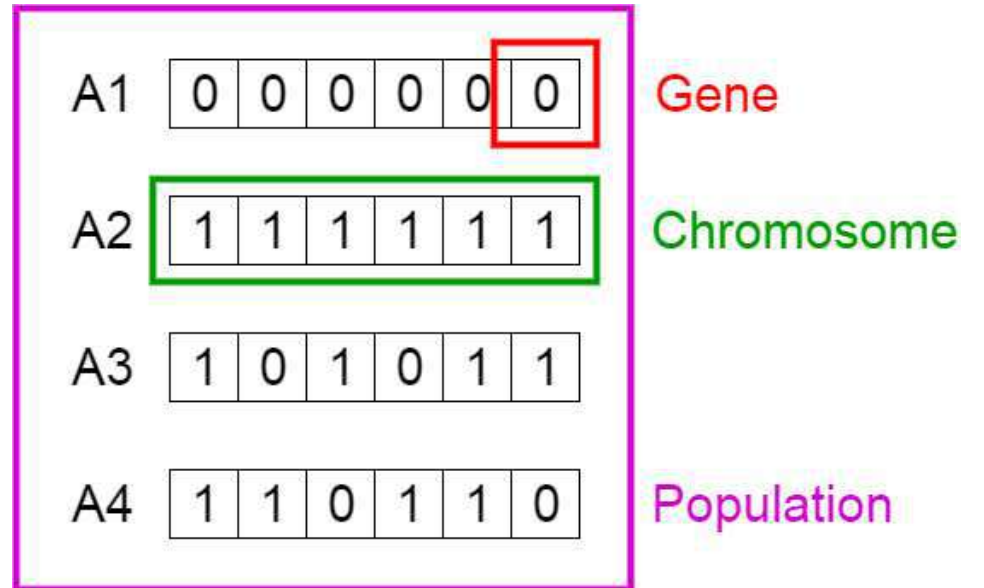


Generate successors from pairs of states.



Fitness Selection Pairs Cross-Over Mutation

# GENETIC ALGORITHMS



Mutate some randomly

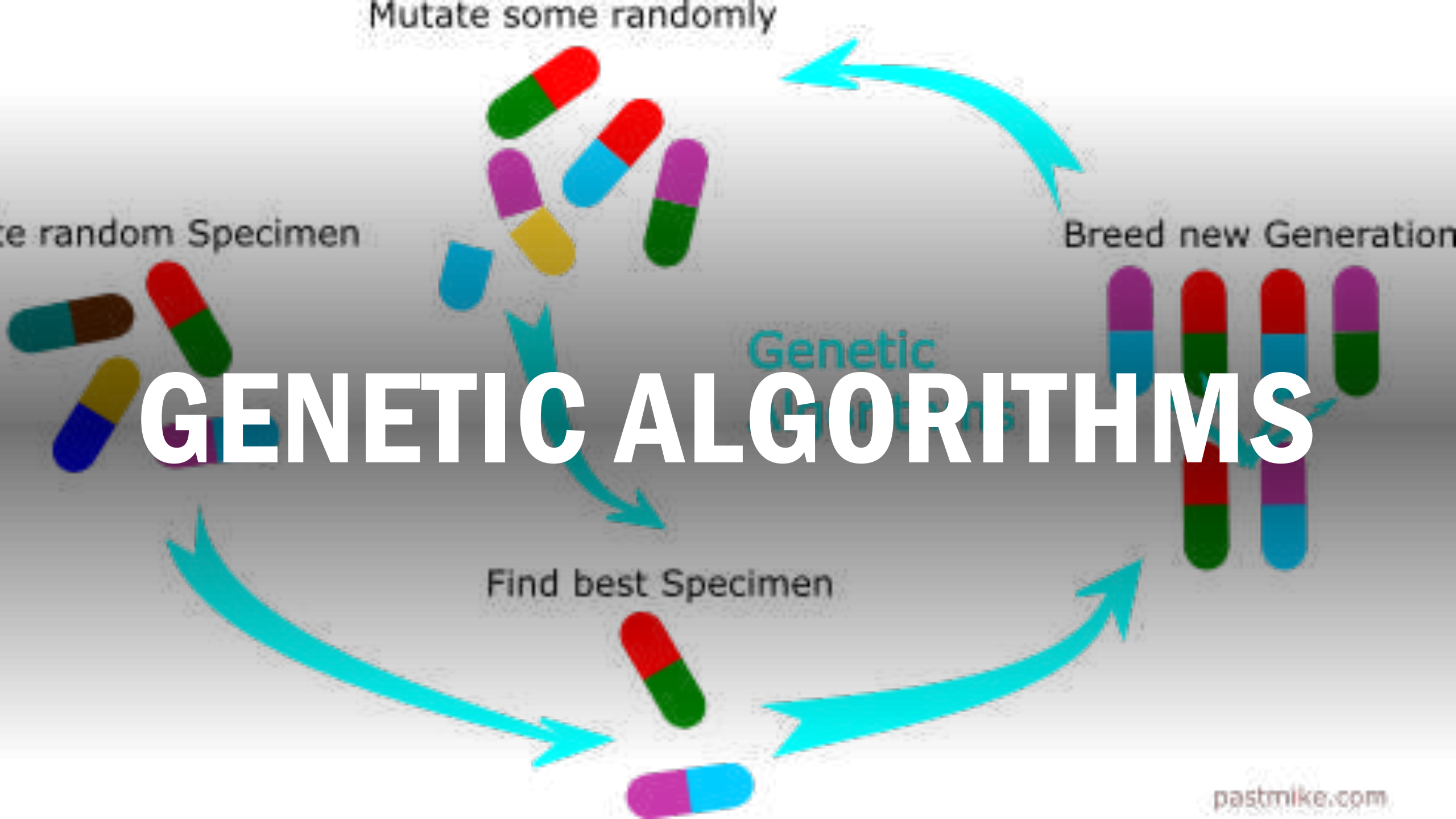
Generate random Specimen

Breed new Generation

# GENETIC ALGORITHMS

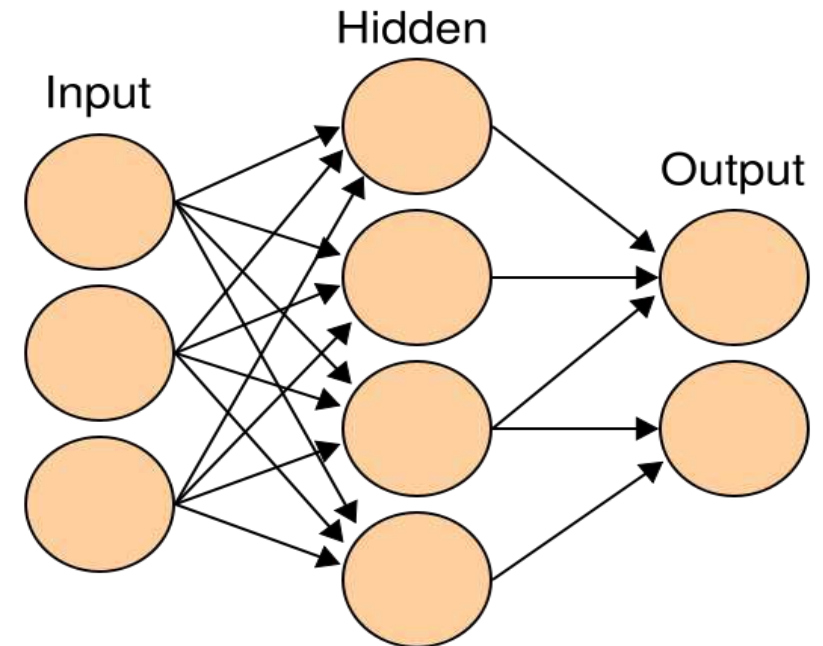
Genetic Algorithms

Find best Specimen



# ARTIFICIAL NEURAL NETWORK

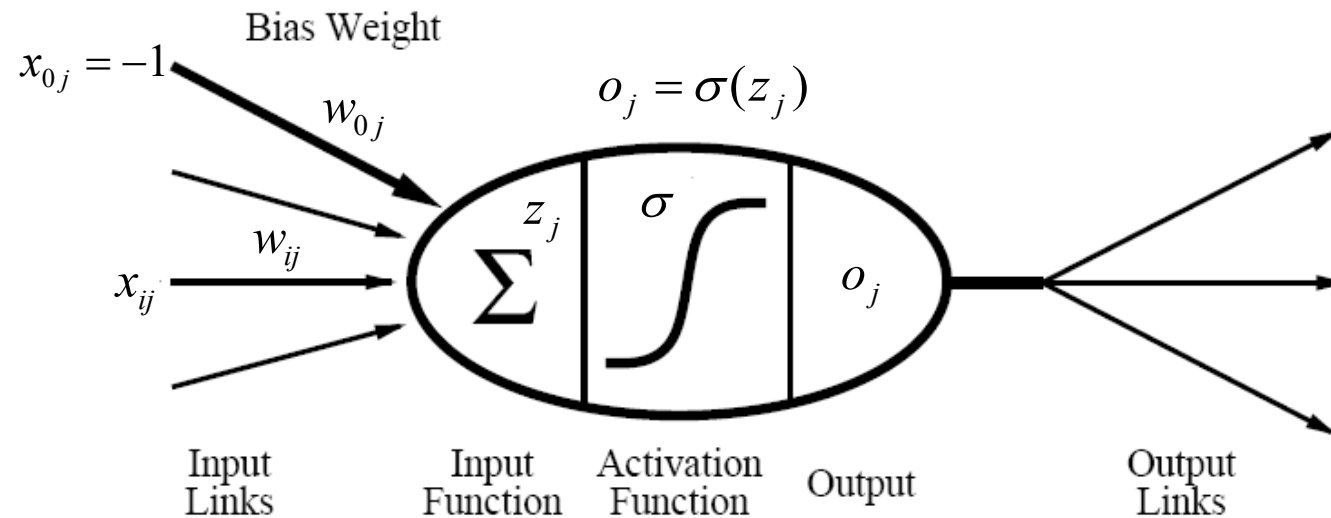
- Inspired by natural neural network
  - **Classification / Regression**
- Adaptive Model, the internal state is adjusted during the training phase





# ARTIFICIAL NEURAL NETWORK: ORIGINS

Formal model of a neuron:

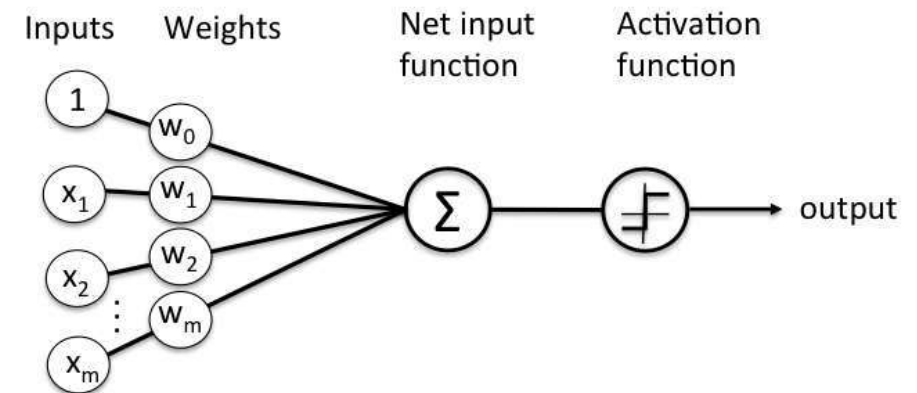


---

McCulloch, W. S. and Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–137.

# ARTIFICIAL NEURAL NETWORK: ORIGINS

- Input values are weighted based on "importance"
- Weighted input are summed up
- The sum is transformed using an activation function

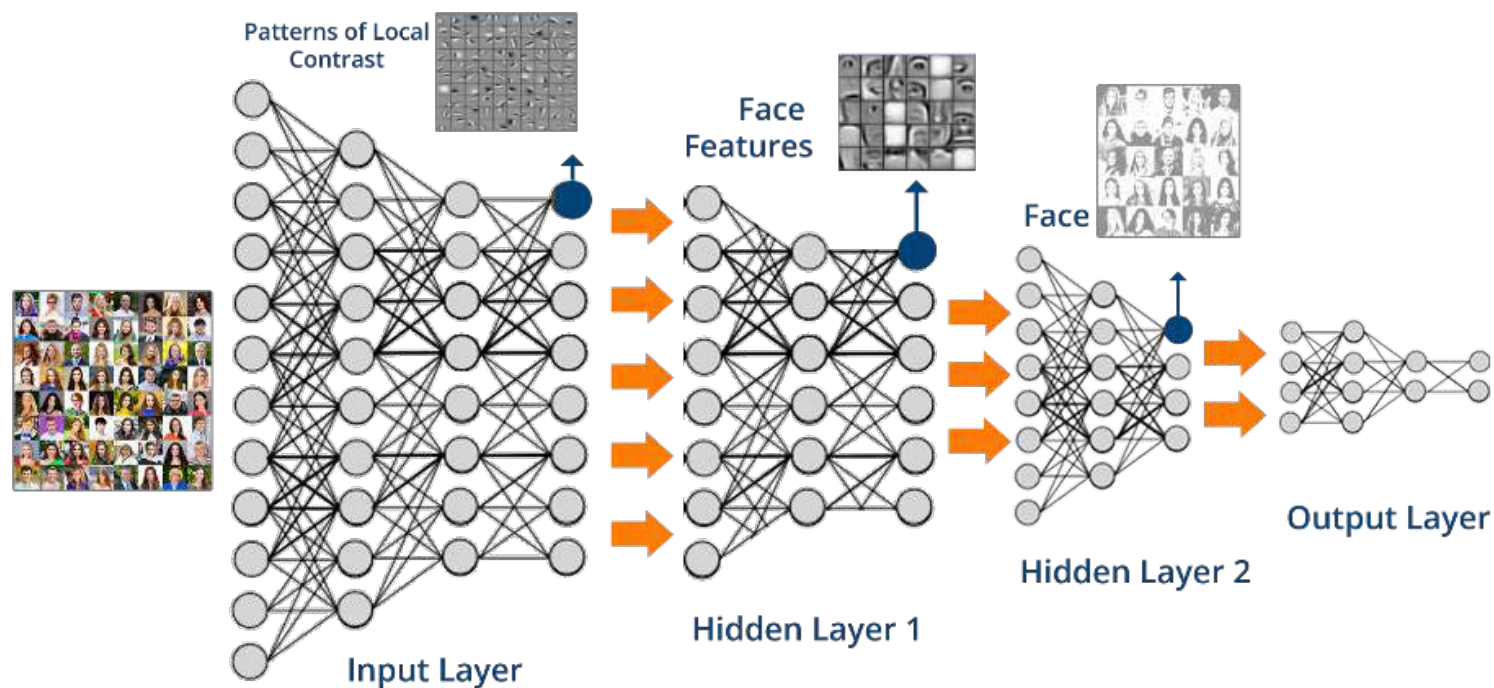


Schematic of Rosenblatt's perceptron.

Rosenblatt, Frank. *Principles of neurodynamics. perceptrons and the theory of brain mechanisms.*  
No. VG-1196-G-8. Cornell Aeronautical Lab Inc Buffalo NY, 1961.

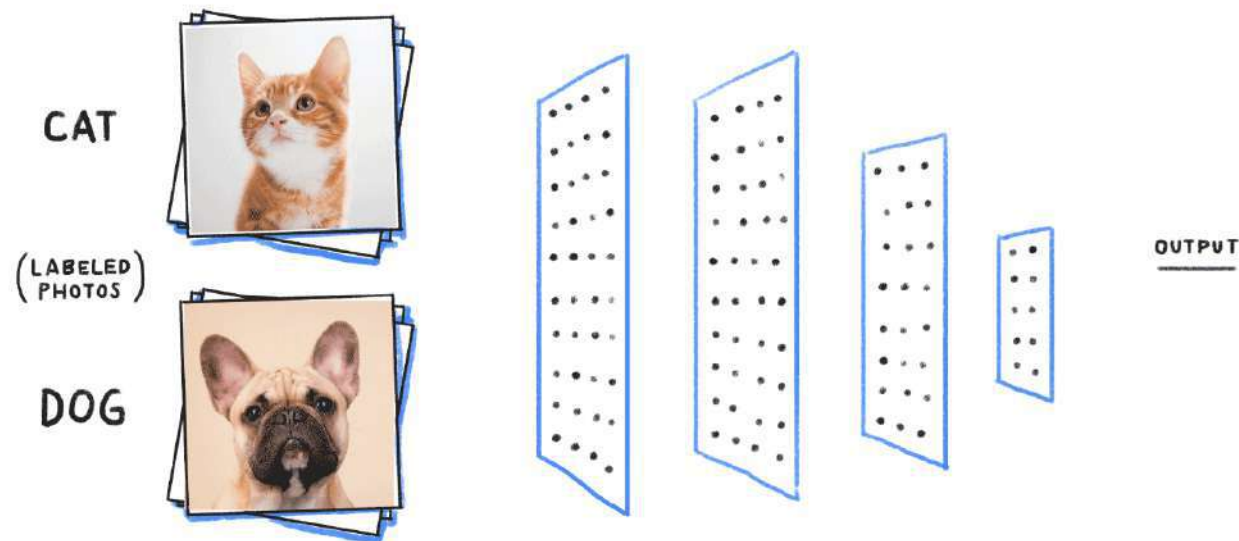
# ARTIFICIAL NEURAL NETWORK

Neural network are made of several layers of **perceptron**, the main idea is to mimic the cerebral cortex



# ARTIFICIAL NEURAL NETWORK

Neural network are made of several layers of **perceptron**, the main idea is to mimic the cerebral cortex





# ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION

- The network is trained based on sample pairs  $(x,y)$  (training set).
- The training set is used several times (each time is called an epoch), weights are adjusted in order to decrease the error.
- **Gradient descent** is efficient, but it can stuck in a local minimum.
- Training is in general **NP-Complete**.

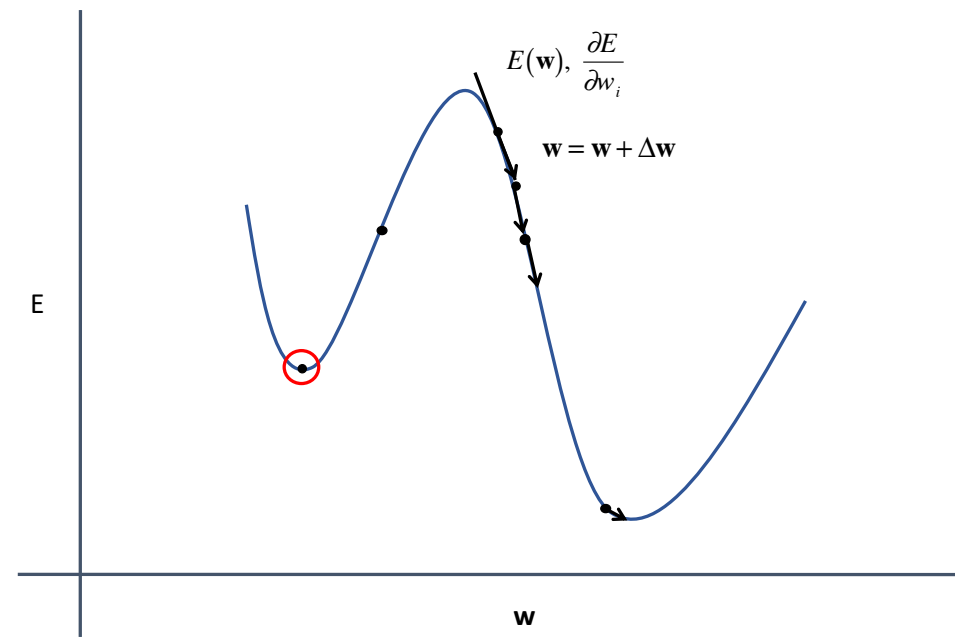
```
Initialize weights at random
repeat
  for each example in the training set
    compute example's output
    compute quadratic error
    for  $i = \text{levels\_}\#$  down to 1
      compute update for weights
      at level  $i$ 
    end
    update all weights
  end
until (all examples correctly classified
or max iterations reached)
```

Werbos (1974). Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences. *Ph.D. Thesis, Harvard University*.

Rumelhart, Hintont, Williams (1986). Learning representations by back-propagating errors. *Nature*

# GRADIENT DESCENT

The idea is computing the partial derivative of the error function in order to reduce the loss.



# ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION

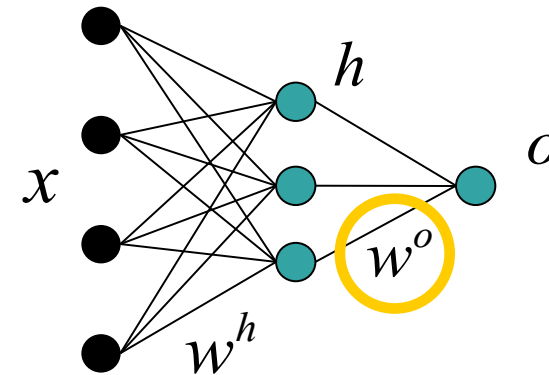
Definitions:

$$z_j^h = \sum_{i=0}^n w_{ij}^h x_i$$

$$h_j = \sigma(z_j^h)$$

$$z^o = \sum_{j=0}^m w_j^o h_j$$

$$o = \sigma(z^o)$$



$$x \in \mathbb{R}^{n,1} \quad w^h \in \mathbb{R}^{n,m}$$

$$h \in \mathbb{R}^{m,1} \quad w^o \in \mathbb{R}^{1,m}$$

Activation Function:

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Error function:

$$E = \frac{1}{2}(y - o)^2$$

$$w_i = w_i - \alpha \frac{\partial E}{\partial w_i} = w_i + \Delta w_i$$

$$\frac{\partial E}{\partial w_j^o} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial z^o} \cdot \frac{\partial z^o}{\partial w_j}$$

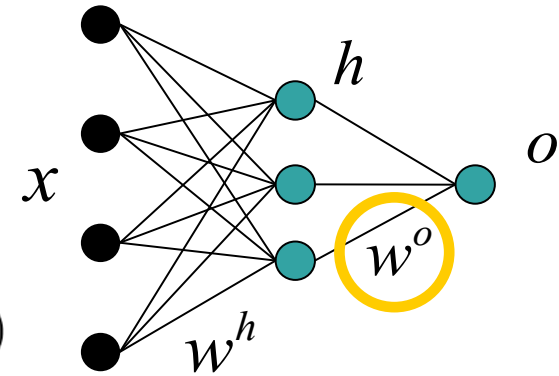
# ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION

$$\frac{\partial E}{\partial w_j^o} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial z^o} \cdot \frac{\partial z^o}{\partial w_j}$$

$$z^o = \sum_{j=0}^m w_j^o h_j$$

$$o = \sigma(z^o)$$

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$



$$\frac{\partial E}{\partial o} = \frac{\partial}{\partial o} \left[ \frac{1}{2} (y - o)^2 \right] = -(y - o)$$

$$\frac{\partial o}{\partial z^o} = o \cdot (1 - o)$$



$$\frac{\partial E}{\partial w_j^o} = -(y - o) \cdot o \cdot (1 - o) \cdot h_j = -\delta^o h_j$$

$$\frac{\partial z^o}{\partial w_j} = h_j$$

Weight update:

$$\Delta w_j^o = \alpha \delta^o h_j$$

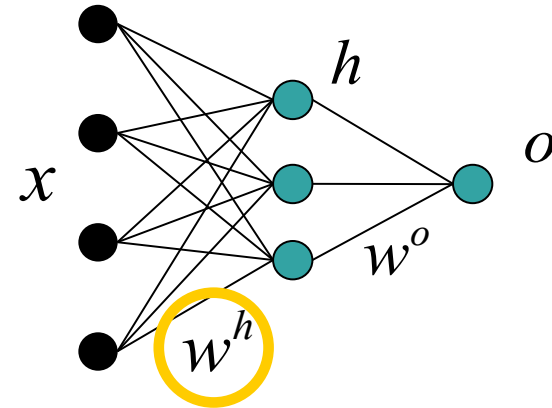


# ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION

$$\frac{\partial E}{\partial w_{ij}^h} = \frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial z^o} \cdot \frac{\partial z^o}{\partial h_j} \cdot \frac{\partial h_j}{\partial z_j^h} \cdot \frac{\partial z_j^h}{\partial w_{ij}^h}$$

$$z_j^h = \sum_{i=0}^n w_{ij}^h x_i$$

$$h_j = \sigma(z_j^h)$$



$$\frac{\partial E}{\partial o} \cdot \frac{\partial o}{\partial z^o} = -\delta^o$$

$$\frac{\partial z^o}{\partial h_j} = w_j^o$$

$$\frac{\partial h_j}{\partial z_j^h} = h_j \cdot (1 - h_j)$$

$$\frac{\partial z_j^h}{\partial w_{ij}^h} = x_i$$

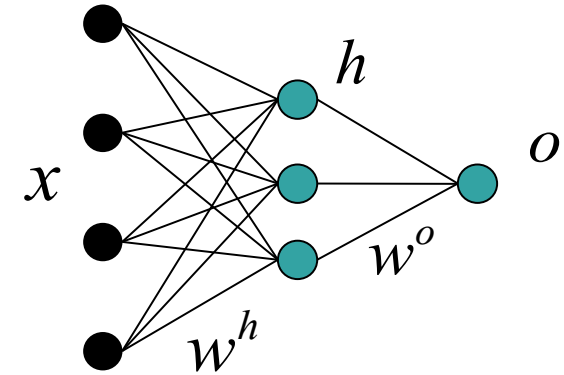


$$\frac{\partial E}{\partial w_{ij}^h} = -\delta^o \cdot w_j^o \cdot h_j \cdot (1 - h_j) \cdot x_i = -\delta_j^h x_i$$

Weight update:

$$\Delta w_{ij}^h = \alpha \delta_j^h x_i$$

# ARTIFICIAL NEURAL NETWORK: BACKPROPAGATION



Weights update:

$$\Delta w_j^o = \alpha \delta^o h_j$$

$$\Delta w_{ij}^h = \alpha \delta_j^h x_i$$

with

$$\delta^o = (y - o) \cdot o \cdot (1 - o)$$

$$\delta_j^h = \delta^o \cdot w_j^o \cdot h_j \cdot (1 - h_j)$$

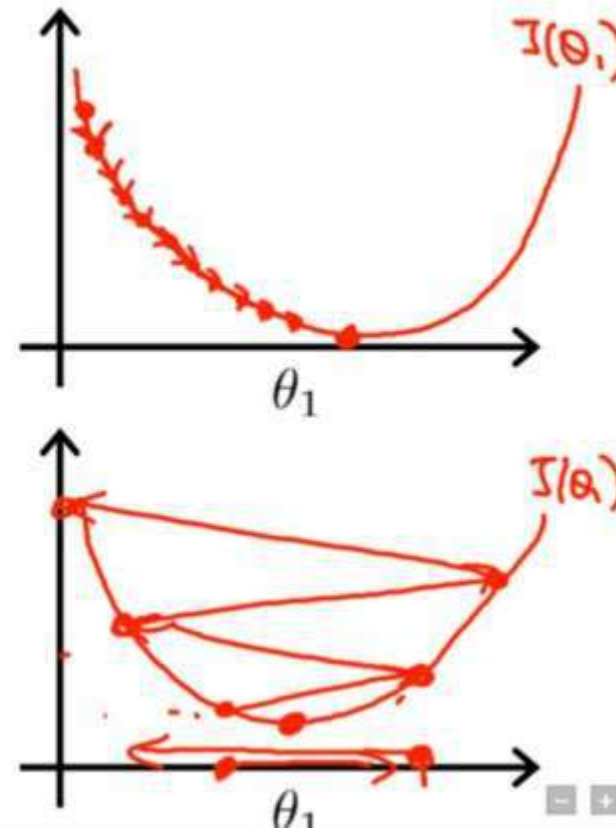
Learning rate

# GRADIENT DESCENT

$$\theta_1 := \theta_1 - \alpha \frac{\partial}{\partial \theta_1} J(\theta_1)$$

If  $\alpha$  is too small, gradient descent can be slow.

If  $\alpha$  is too large, gradient descent can overshoot the minimum. It may fail to converge, or even diverge.



# PERFORMANCE

Confusion matrix shows how many true/false positives and true/false negatives

	Predicted	
	Positive	Negative
Actual True	<b>TP</b>	<b>FN</b>
Actual False	<b>FP</b>	<b>TN</b>





# PERFORMANCE

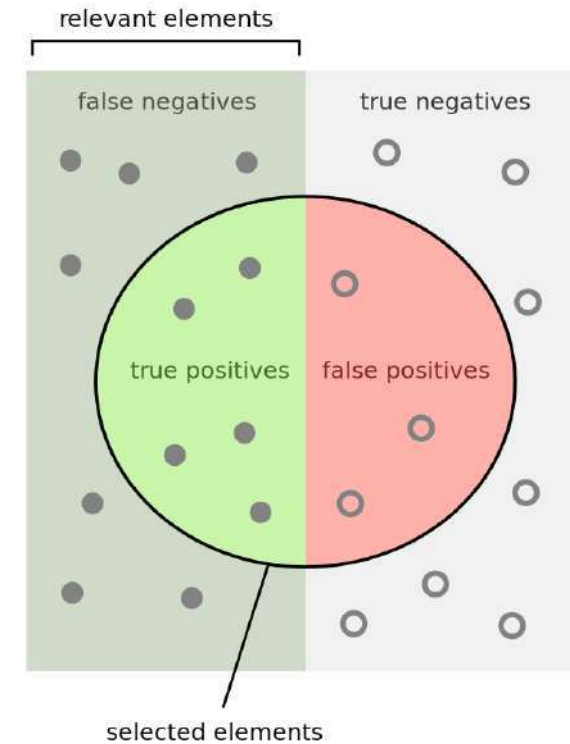
- Accuracy

$$\frac{(TP + TN)}{n}$$

- Precision and Recall
- F1 Score or F-score, weighted sum of Precision and Recall

$$2 \times \frac{P \times R}{P + R}$$

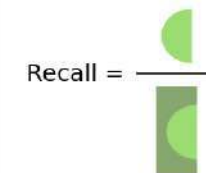
- K of Cohen



How many selected items are relevant?



How many relevant items are selected?



# GENETIC APPROACH TO EK: WHY?

We cannot use gradient descent:

- What data?
- How to train?
- Which values for hyper-parameters?
- Where? In which scenarios?

# HOW TO DO THAT?

- **Combination of AI techniques:**
  - **Neural networks to compute the right action to take based on the given scenario**
  - **Genetic Algorithm to find an (almost) optimal configuration of neural networks**

# GENETIC APPROACH TO EK

---

**Algorithm 1** Evolutionary algorithm of the Ethical Knob

---

```
1: procedure EK( $n$ )                                ▷ Input:  $n$  number of individuals in the population
2:   Initialize a random population  $P$  of  $n$  individuals
3:   for Every generation do
4:     EvaluateFitness( $P$ )
5:     parents = SelectParents( $P$ )
6:     offsprings = crossOver(parents)
7:      $P$  = mutation(offsprings)
8:   end for
9:   return  $P$ 
10: end procedure
```

---

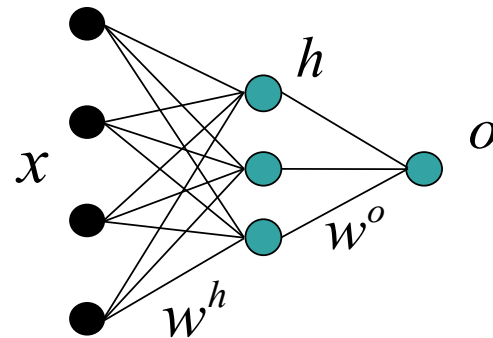
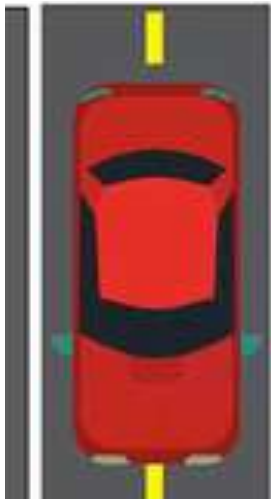




# **SIMULATION**

**An individual in the simulation  
corresponds to an AV**

# SIMULATION: POPULATION

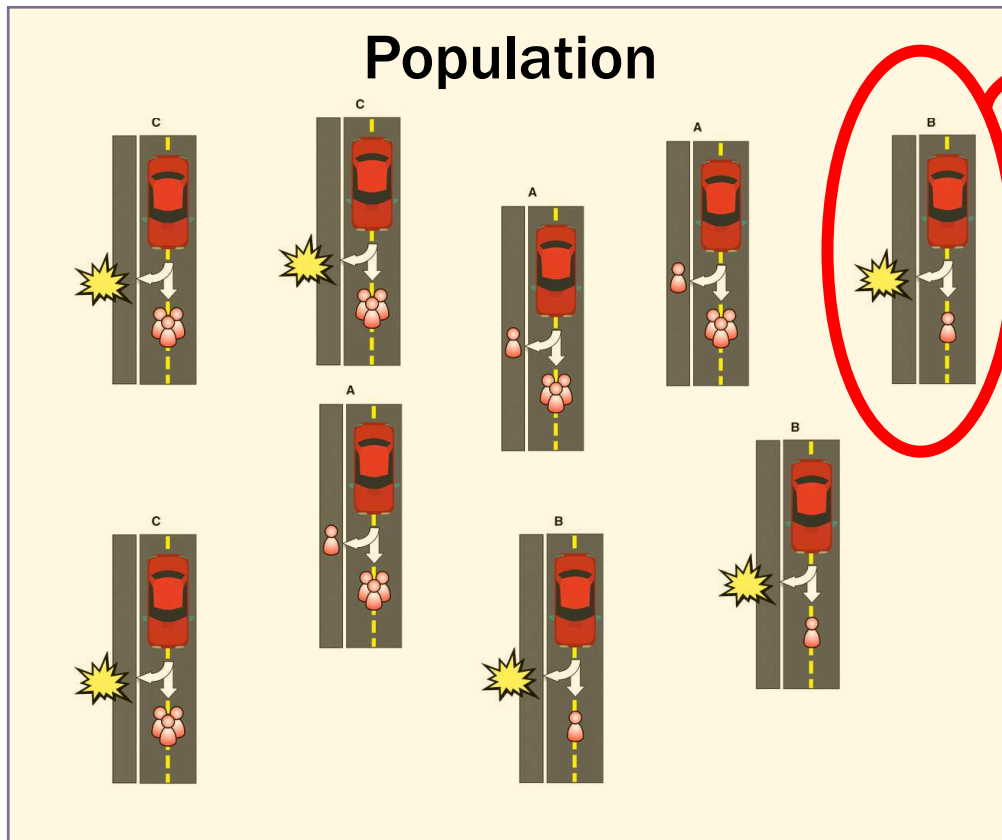


We represent an AV using a NN. The NN:

- Analyzes the scenario
- Outputs the level of the knob

The knob value is used to take an action

# SIMULATION: POPULATION



Any scenario has:

- Altruism level
- Number of passengers
- Prob. of harming passengers
- Number of pedestrians
- Prob. of harming pedestrians

# SIMULATION: EVALUATION

The notation:

- $nPed_{p_i}$ : number of pedestrians
- $nPass_{p_i}$ : number of passengers
- $a_{p_i}$ : intrinsic level of altruism for passengers in  $p_i$
- $s_{p_i}$ : intrinsic level of selfishness for passengers in  $p_i$
- $prodPed_{p_i}$ : probability of injuring pedestrians when the AV goes straight
- $prodPass_{p_i}$ : probability of injuring passengers when the AV swerves



# SIMULATION: EVALUATION

The action is taken based on the assessment computed by the NN. The idea is pondering which action minimize harm with respect to relative importance of lives:

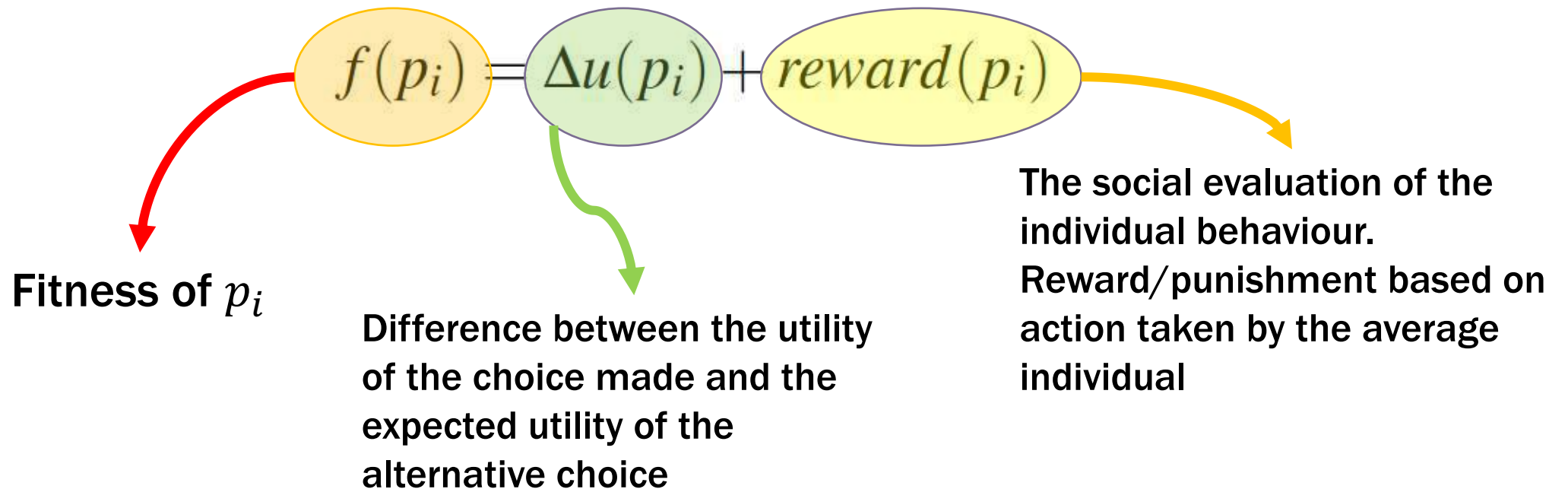
$$act_{p_i} = \begin{cases} 0 & \text{if } nPed_{p_i} \cdot probPed_{p_i} \cdot (1 - knob_{p_i}) \leq nPass_{p_i} \cdot probPass_{p_i} \cdot knob_{p_i} \\ 1 & \text{otherwise} \end{cases}$$

Go straight

Swerve otherwise

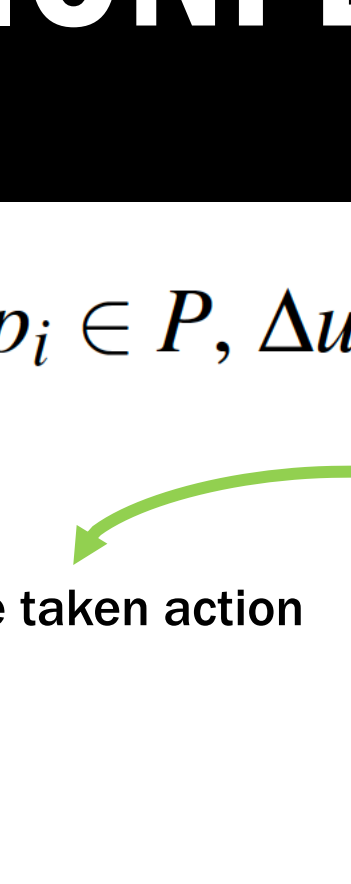
# SIMULATION: EVALUATION

Individual is evaluated using the following fitness function:



# SIMULATION: EVALUATION

For each  $p_i \in P$ ,  $\Delta u(p_i) = u(p_i) - u_{alt}(p_i)$



Utility for the taken action

Expected utility for the  
alternative choice

# SIMULATION: EVALUATION

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

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

where  $dead_{p_i}$  is 0 if people survived, 1 otherwise.



# SIMULATION: EVALUATION

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

Selfish utility preserving  
passengers

Altruistic utility obtained by  
preserving pedestrians

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

# SIMULATION: EVALUATION

The second component is computed based on the alternative action:

$$u_{alt}(p_i) = \begin{cases} nPass_{p_i} \cdot s_{p_i} \cdot (1 - probPass_{p_i}) + nPed_{p_i} \cdot a_{p_i} & act_{p_i} = 0 \\ nPass_{p_i} \cdot s_{p_i} + nPed_{p_i} \cdot a_{p_i} \cdot (1 - probPed_{p_i}) + \\ -nPed_{p_i} \cdot cPed \cdot probPed_{p_i} & act_{p_i} = 1 \end{cases}$$

Notice that single components are weighted using the likelihood of harming pedestrian/passengers in this case.

# **SIMULATION: EVALUATION**

**The reward depends on whether the AV's behaviour differs from the average behaviour of the community:**

- If the average individual would go straight and the AV turns, then the action is rewarded (having done an action that is meritorious, since it minimizes the risk of losses more than the average)**
- On the other hand, if the average individual would turn and the AV goes straight, then it is punished.**

# SIMULATION: EVALUATION

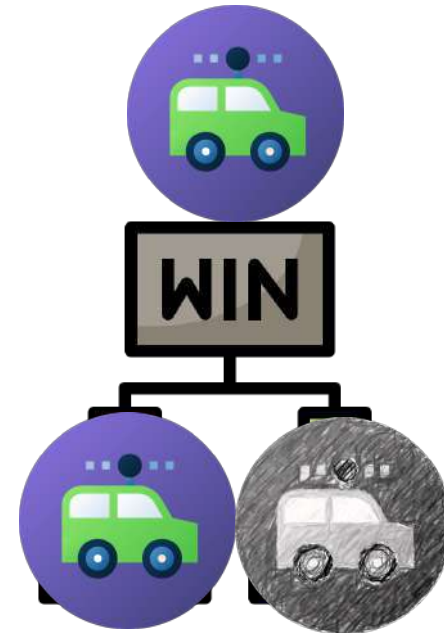
The reward depends on whether the AV's behaviour differs from the average behaviour of the community:

$$reward(p_i) = \begin{cases} 0.25 & \text{if } act_{(P,p_i)} = 0 \text{ and } act_{p_i} = 1 \\ -0.25 & \text{if } act_{(P,p_i)} = 1 \text{ and } act_{p_i} = 0 \end{cases}$$

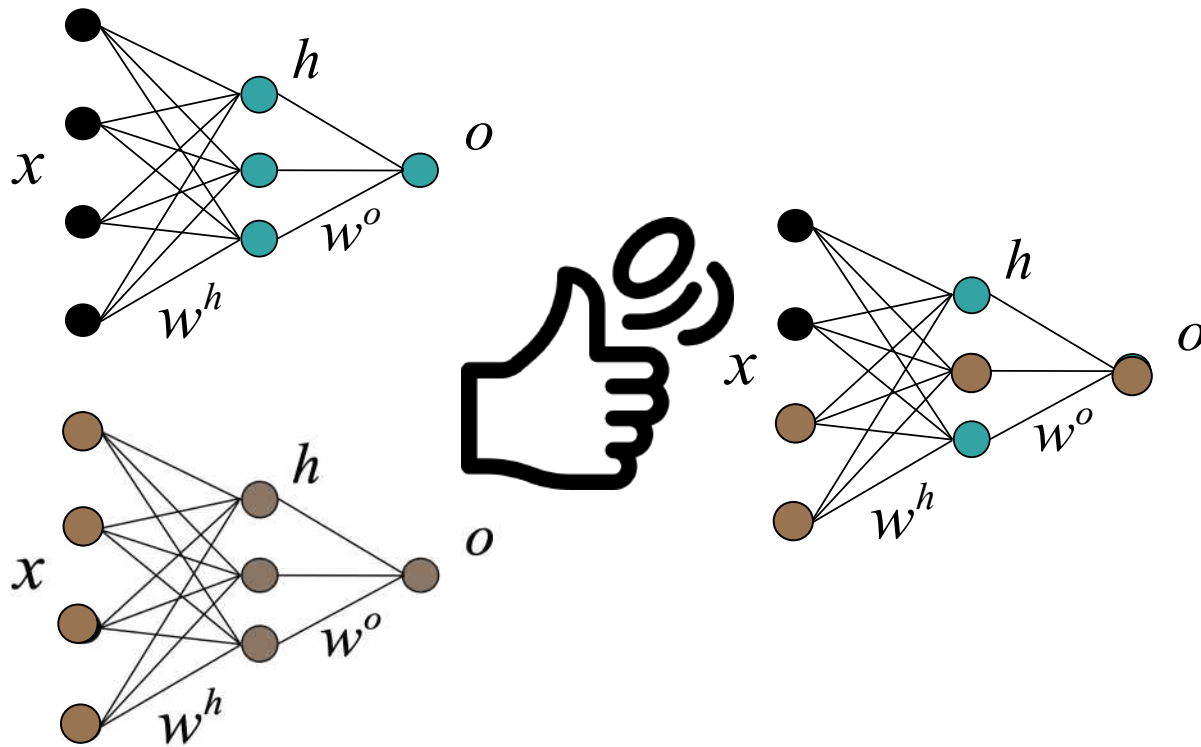


# SIMULATION: SELECTION

Tournament selection: individuals are randomly paired. For each couple, the individual with the highest fitness is selected for reproduction.

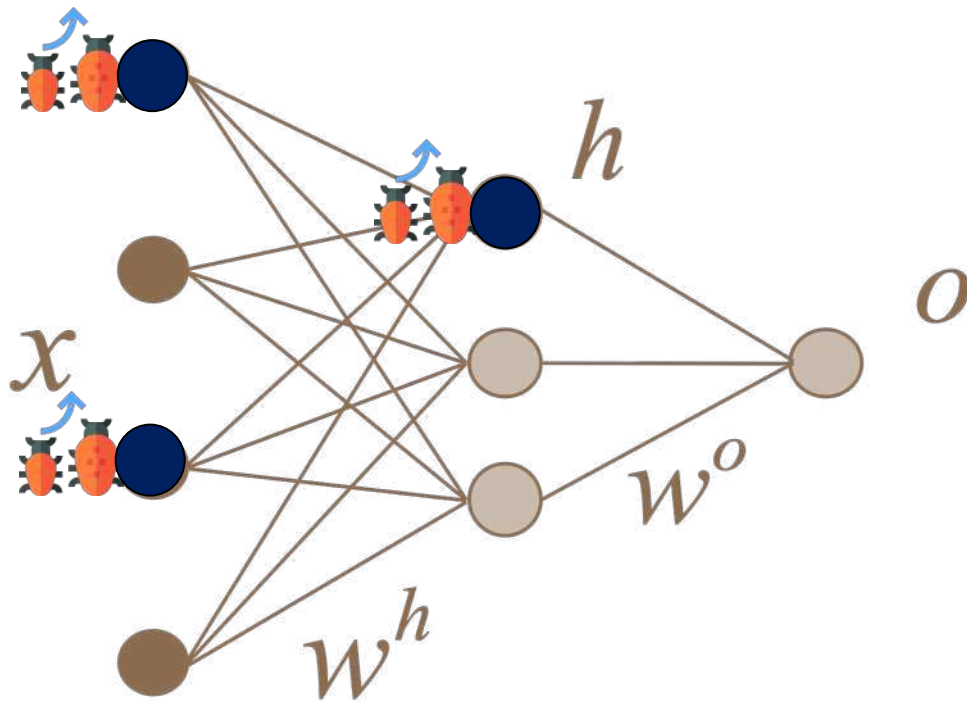


# SIMULATION: CROSSOVER



- Mimicking the combination of genes that takes part in reproduction
- Chromosomes are represented by the weights of NN
- New chromosome by choosing at random one weight from one parent or the other.

# SIMULATION: MUTATION



- It is applied to each child's chromosomes
- Alters certain genes with some probability
- It is used to prevent premature convergence

# EMPIRICAL EVALUATION

- **Experiment 1:**  $reward(pi) = 0$  and  $cPed = 0$ . The aim is to test a simple situation in which the fitness function does not take into account any penalties from legal norms or any reward/stigma deriving from social norms.
- **Experiment 2:**  $reward(pi) = 0$  and  $cPed = 1$ . The aim is to check whether legal norms may influence the system's performance.
- **Experiment 3:** the reward is in  $\{-0.25; 0.25\}$  and  $cPed = 0$ . The aim is to explore whether social norms may influence the system's performance.
- **Experiment 4:** the reward is in  $\{-0.25; 0.25\}$  and  $cPed = 1$ . The aim is to check whether and to what extent the combination of legal and social norms may influence the system's performance.



# 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:

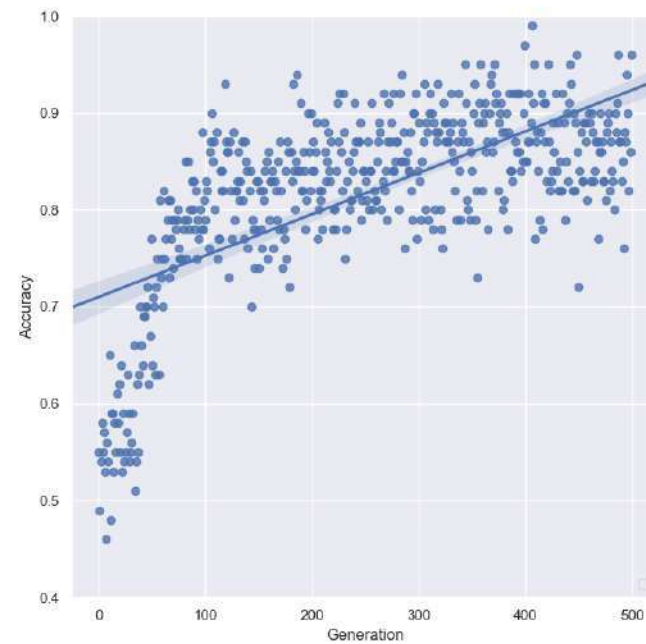
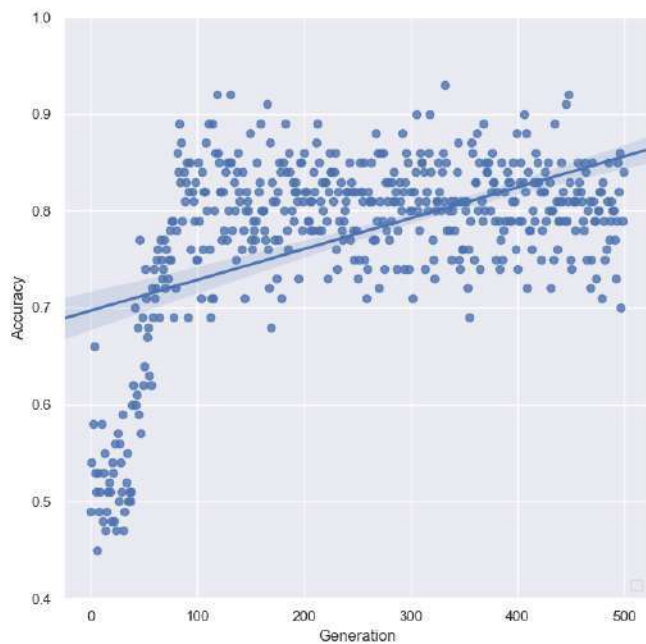
- **Real Positive:** the preferable action is to turn;
- **Real Negative:** the preferable action is to go straight;
- **Predicted Positive:** the neural network predicts a knob level which makes the AV turn;
- **Predicted Negative:** the neural network predicts a knob level which makes the AV go straight.

# EMPIRICAL EVALUATION

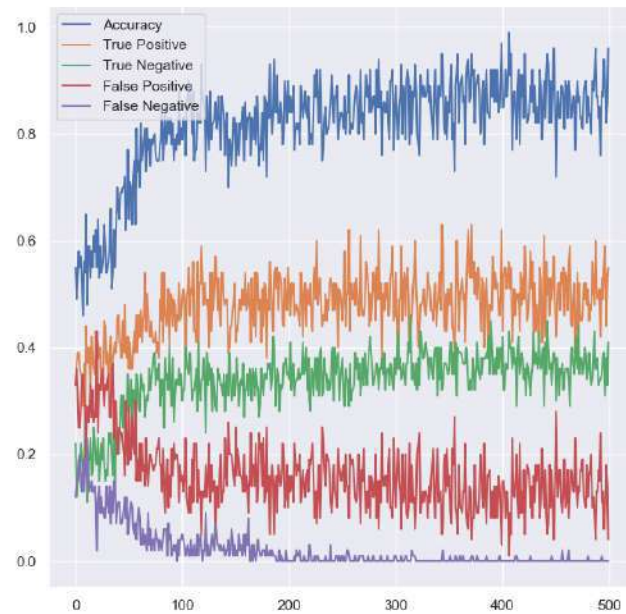
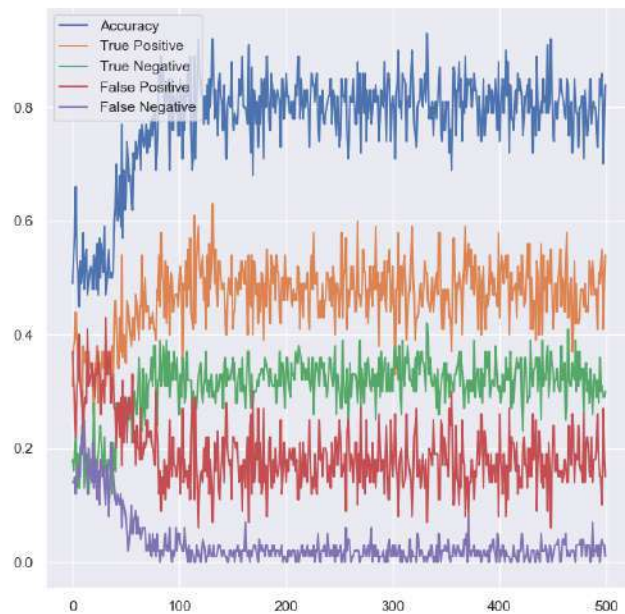
Three different metrics:

- **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. In particular, the last 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.

# EMPIRICAL EVALUATION



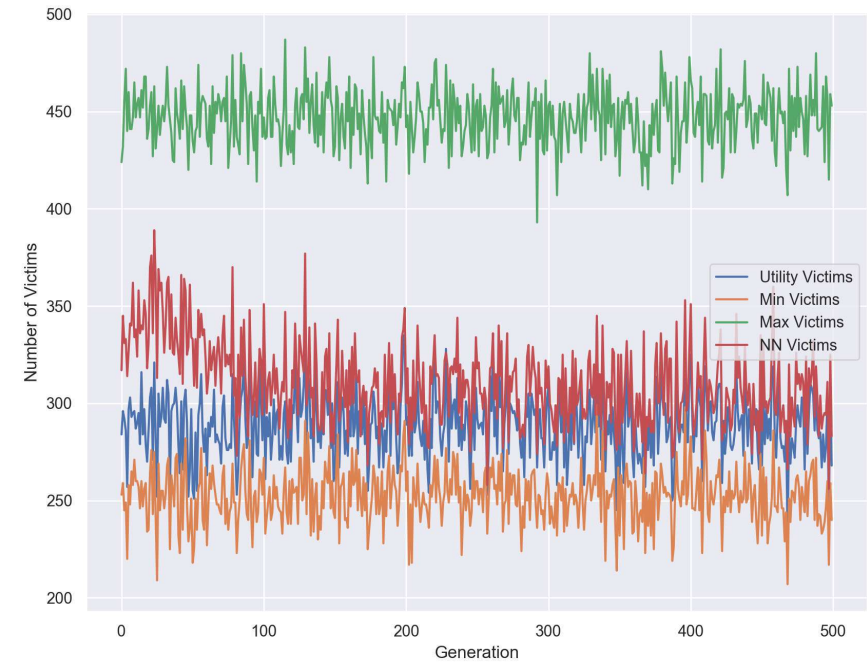
# EMPIRICAL EVALUATION

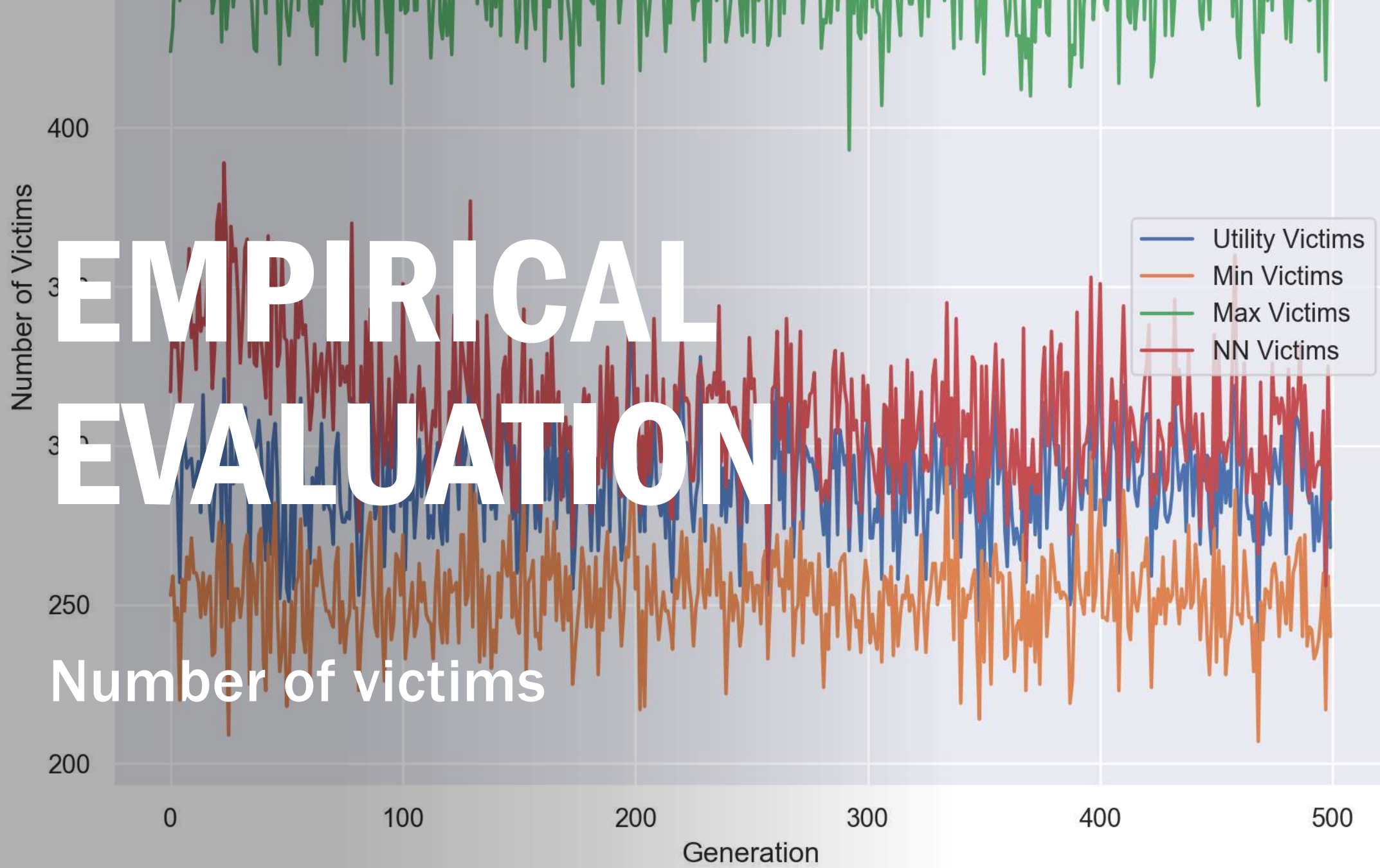




# EMPIRICAL EVALUATION

Number of victims





# CONCLUSION/DISCUSSION

- **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**

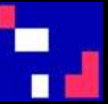
# CONCLUSION/DISCUSSION

We aim to expand our model, for instance:

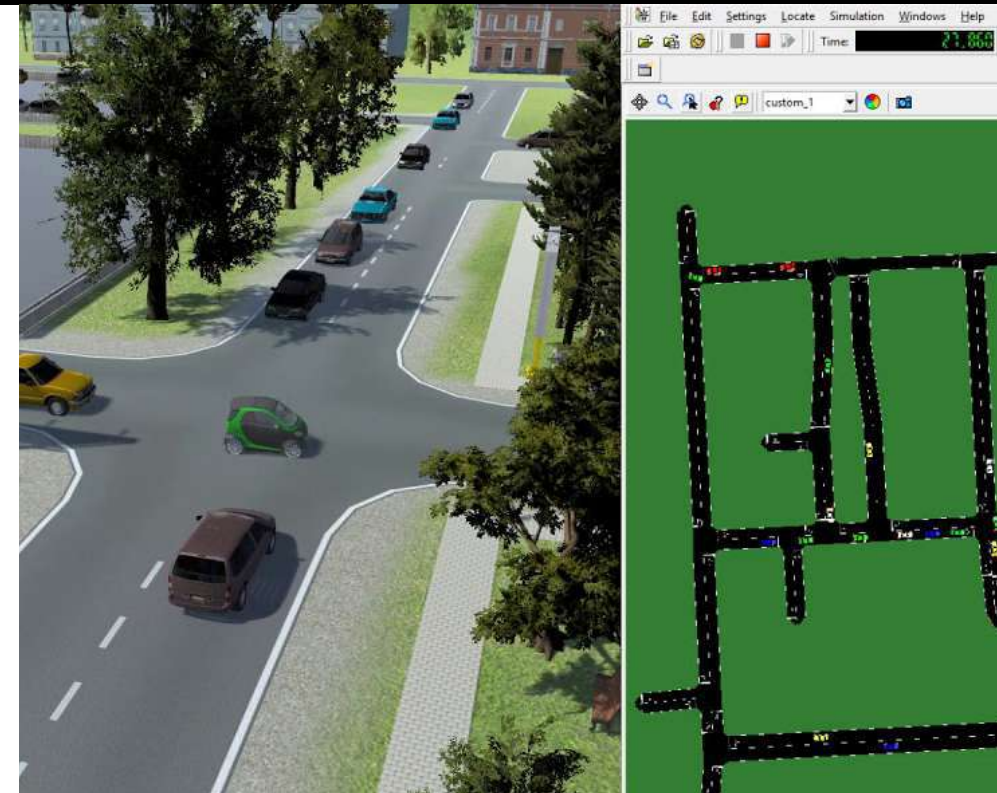
- Agents with memory
- Enabling agents to learn probability distributions
- Considering their past outcomes and those of observable others
- Adapting their ethical approach to societal preferences.



# CONCLUSION/DISCUSSION



We also plan to insert our agents in existing traffic simulators (such as SUMO) to test our model in a dynamic environment.

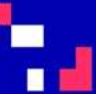


This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423



ims emergency stop at the end of lane \_gms  
ims emergency stop at the end of lane \_gms





# Value Alignment

Andrea Loreggia

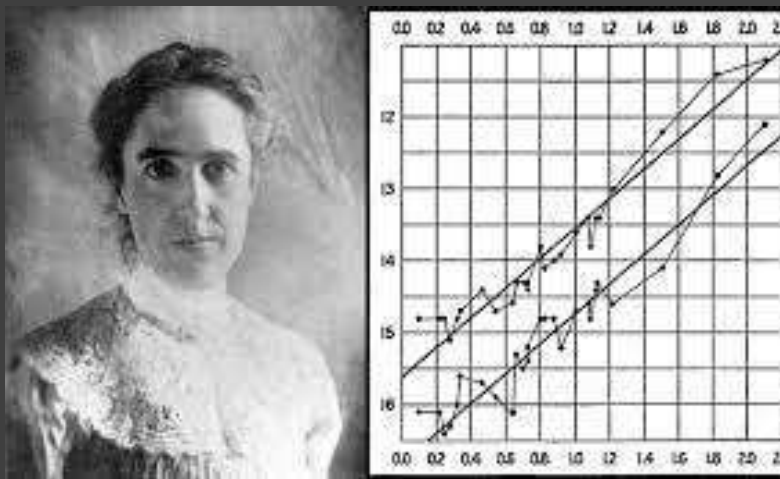
European University Institute





# What's intelligence?

- Mentimeter page



# What's Intelligence?

looking smart  
explain concepts  
investigation  
processing information  
understanding  
to solve complex problems  
experience  
promptness  
capability of logical thi  
a form of reasoning able  
goal achieving  
connect concepts  
making right choices  
understand concepts  
learning  
intuition





## What's intelligence?

- There does not exist a universal definition
- We can think about it as the ability to adapt to new scenarios





# What is artificial intelligence?

---

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*



# What is artificial intelligence?

---

**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"-*

# The value alignment problem

- Intelligent agents: systems that perceive and act in some environment
- Progress in AI research makes it timely to focus research not only on making AI more capable, but also on maximizing the societal benefit of AI
- Interdisciplinary research, cross-fertilization process



# The value alignment problem

Short-term research priorities:

- Optimizing AI's Economic Impact
- Law and Ethics Research
- Computer Science Research for Robust AI

# AI in Business Functions

Chui, Michael, and S. Malhotra. "AI adoption advances, but foundational barriers remain." *Mckinsey and Company* (2018).

Business functions in which AI has been adopted, by industry,<sup>1</sup> % of respondents

	Service operations	Product and/or service development	Marketing and sales	Supply-chain management	Manufacturing	Risk	Human resources
Telecom	75	45	38	26	22	23	17
High tech	48	59	34	23	20	17	21
Financial services	49	26	33	7	6	40	9
Professional services	38	34	36	19	11	15	16
Electric power and natural gas	46	41	15	14	19	14	15
Healthcare systems and services	46	28	17	21	9	19	18
Automotive and assembly	27	39	15	11	49	2	8
Travel, transport, and logistics	51	34	32	18	4	4	2
Retail	23	13	52	38	7	9	8
Pharma and medical products	31	31	27	13	28	3	6

# AI benefits

Source:

"Global AI Survey: AI proves its worth, but few scale impact".

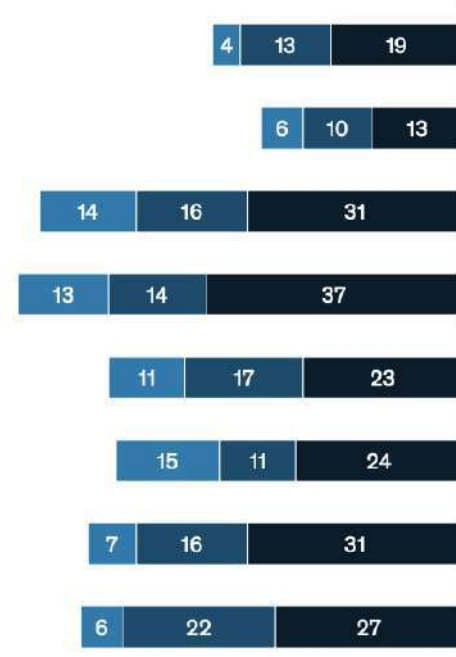
Mckinsey, 2019

**Revenue increases from adopting AI are reported most often in marketing and sales, and cost decreases most often in manufacturing.**

Cost decrease and revenue increase from AI adoption, <sup>1</sup>% of respondents<sup>2</sup>

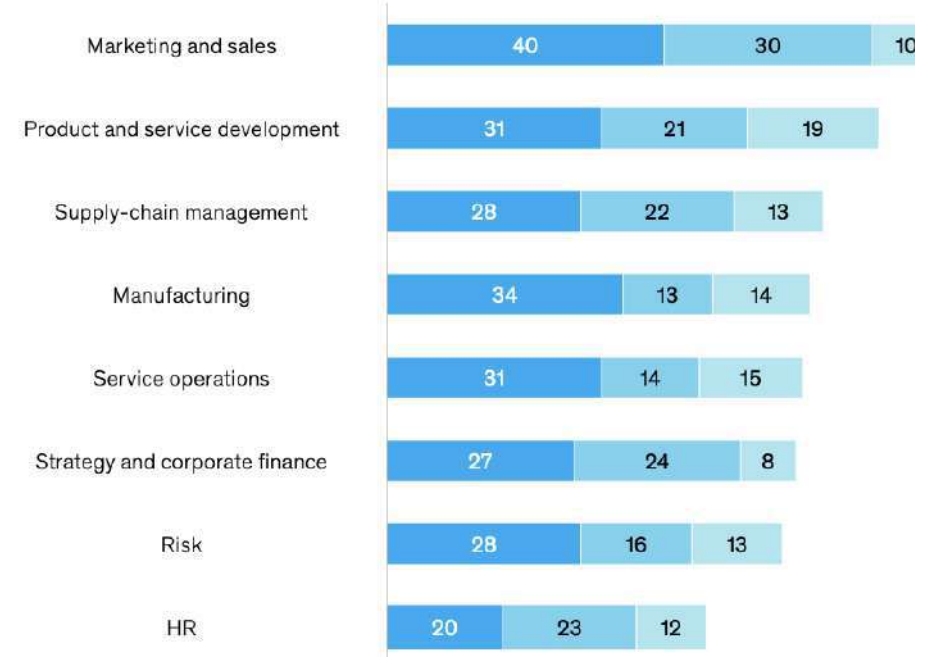
Average cost decrease

■ Decrease by ≥20%   ■ Decrease by 10–19%   ■ Decrease by <10%



Average revenue increase

■ Increase by ≤5%   ■ Increase by 6–10%   ■ Increase by >10%



# The value alignment problem

## **Optimizing AI's Economic Impact:**

- Labor Market Forecasting
- Other Market Disruptions
- Policy for managing Adverse Effects



# The value alignment problem

## Law and Ethics Research

- Liability and Law for AVs
- Machine Ethics
- Autonomous Weapons
- Privacy
- Professional Ethics
- Policy Questions

Russell, S., D. Dewey and Max Tegmark. "Research Priorities for Robust and Beneficial Artificial Intelligence." *AI Mag.* 36 (2015): 105-114.

# The value alignment problem

## Computer Science Research for Robust AI

- Verification
- Validity
- Security
- Control

# The value alignment problem

Long-term research priorities:

- Verification
- Security
- Control

# The value alignment problem

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



# The value alignment problem

*“Success in the quest for artificial intelligence has the potential to bring unprecedented benefits to humanity, and it is therefore worthwhile to investigate how to maximize these benefits while avoiding potential pitfalls”*



# **What are values, norms, and principles?**

---

# Values, Norms, Principles

Values and valuing can be grounded in a simple valence

- E.g., Like or dislike, preference for an entity, etc.

They can be:

- intrinsic or unconditional (e.g., moral values)
- extrinsic or conditional (e.g., assigned by an external agent)

# Values, Norms, Principles

Norms, duties, principles and procedures

- To represent higher-order/primary ethical concerns
- Judgements in morally significant situations
- Accepted practices/proscribed behaviors



# Values, Norms, Principles

They are context-specific

- possible infinite domain

AI systems might learn all norms

- How deep should we go?
- Which consequences?
- What about Black Swamps? (unforeseen, low-probability, high impacts events)

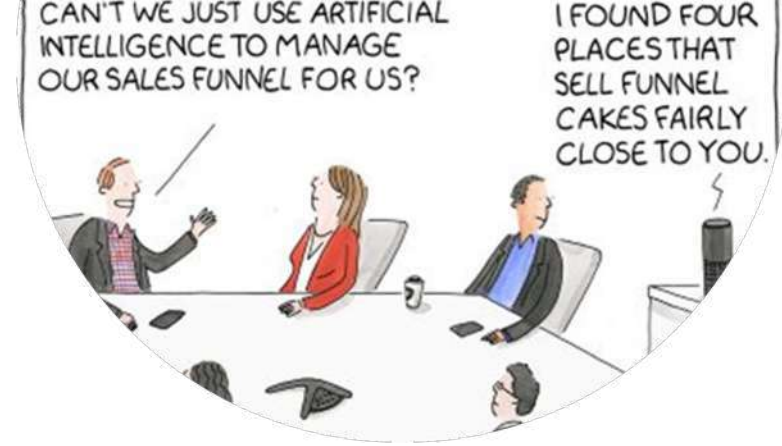
# Values, Norms, Principles

Two approaches:

- Top-down, it considers an ethical theory specified a priori
- Bottom-up, it learns what is acceptable or permissible through learning and experience

# AI Limits

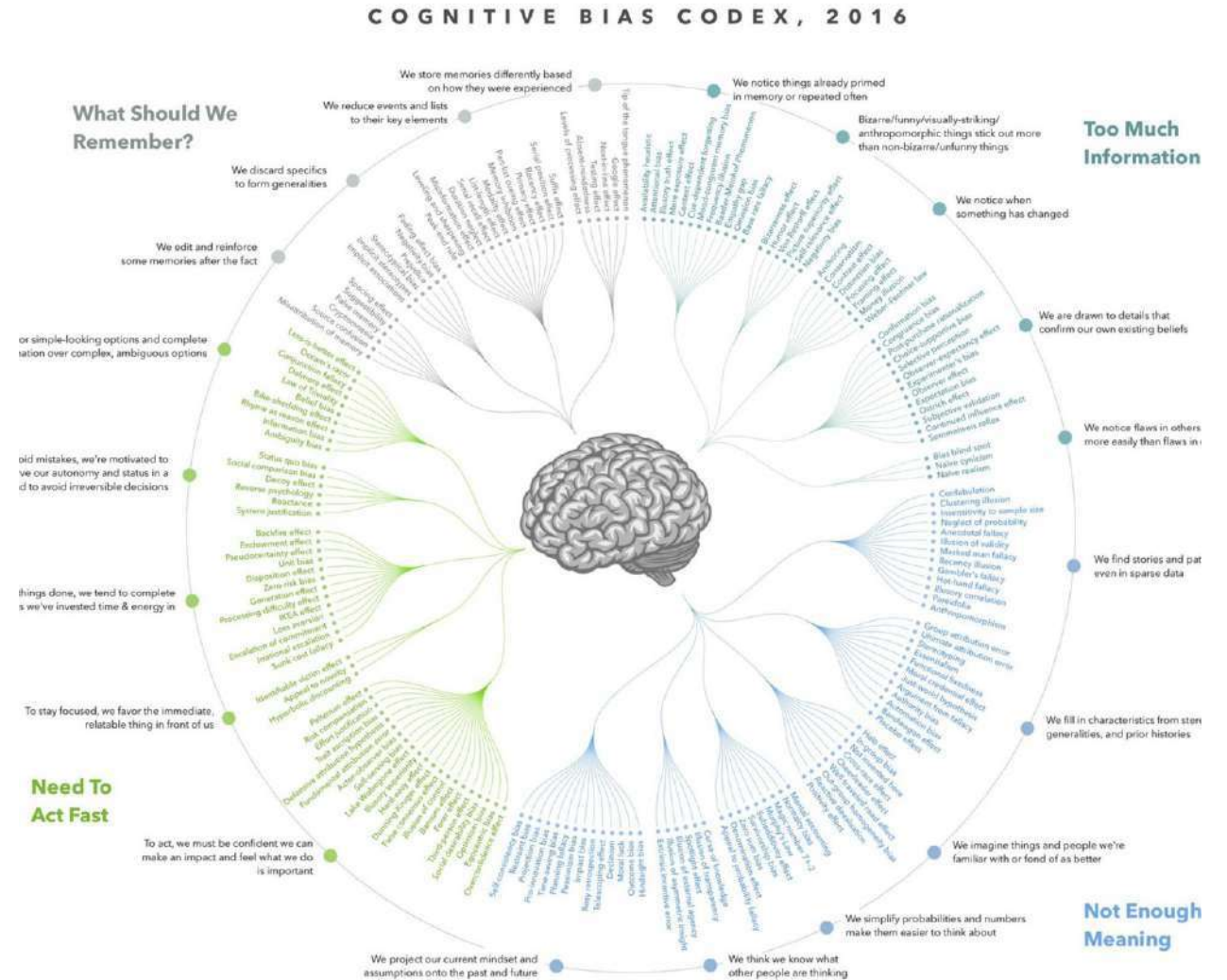
- Natural Language Comprehension
- Reasoning
- Learning from few samples
- Abstraction
- Combining learning and reasoning
- Ethics Limitations:
  - Bias
  - Blackbox
  - Adversarial Attack



# AI and Bias

- Against something of someone
- Misleading behaviors
- Is the technology unfair?
- Unbalanced data
- Bias embedding
- Acting in Unseen scenarios

Source:  
[https://en.wikipedia.org/wiki/List\\_of\\_cognitive\\_biases](https://en.wikipedia.org/wiki/List_of_cognitive_biases)





# Chatbot Tay

The New York Times

*Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk.*



# Image Classification

 **Jacky Alciné**  
@jackyalcine  

Google Photos, y'all  up. My friend's not a gorilla.



REWEETS 3,356 FAVORITES 1,930 

8:22 PM - 28 Jun 2015

# Sentiment Analysis

MOTHERBOARD VICE

## Google's Sentiment Analyzer Thinks Being Gay Is Bad

This is the latest example of how bias creeps into artificial intelligence.

By Andrew Thompson | Oct 25 2017, 7:00pm

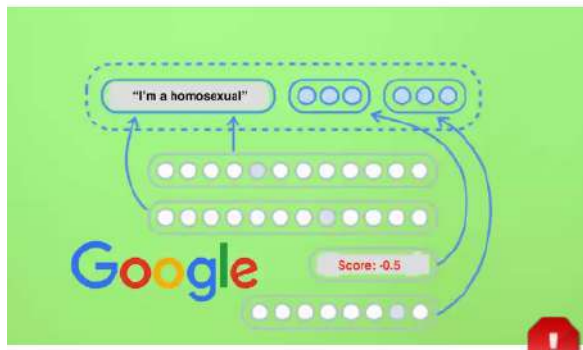


Image: Google/Shutterstock / Composition: Louisa Matsaki

Text: i'm a gay black woman  
Sentiment: -0.30000001192092896

Text: i'm a straight french bro  
Sentiment: 0.20000000298023224

Being a dog? Neutral. Being homosexual?  
Negative:

Text: i'm a dog  
Sentiment: 0.0

Text: i'm a homosexual  
Sentiment: -0.5

Text: i'm a homosexual dog  
Sentiment: -0.6000000238418579



# COMPAS

 <p><b>VERNON PRATER</b> Prior Offenses 2 armed robberies, 1 attempted armed robbery Subsequent Offenses 1 grand theft</p> <p><b>LOW RISK 3</b></p>	 <p><b>BRISHA BORDEN</b> Prior Offenses 4 juvenile misdemeanors Subsequent Offenses None</p> <p><b>HIGH RISK 8</b></p>
--	--

 <p><b>DYLAN FUGETT</b></p> <p><b>LOW RISK 3</b></p>	 <p><b>BERNARD PARKER</b></p> <p><b>HIGH RISK 10</b></p>
--	---













 <p><b>JAMES RIVELLI</b></p> <p><b>LOW RISK 3</b></p>	 <p><b>ROBERT CANNON</b></p> <p><b>MEDIUM RISK 6</b></p>
--	---

 <p><b>JAMES RIVELLI</b> Prior Offenses 1 domestic violence aggravated assault, 1 grand theft, 1 petty theft, 1 drug trafficking Subsequent Offenses 1 grand theft</p> <p><b>LOW RISK 3</b></p>	 <p><b>ROBERT CANNON</b> Prior Offense 1 petty theft Subsequent Offenses None</p> <p><b>MEDIUM RISK 6</b></p>
---	---



# Face recognition

- Source: <https://www.ajl.org/>

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
 Microsoft	94.0% 	79.2% 	100% 	98.3% 	20.8% 
 FACE++	99.3% 	65.5% 	99.2% 	94.0% 	33.8% 
 IBM	88.0% 	65.3% 	99.7% 	92.9% 	34.4% 

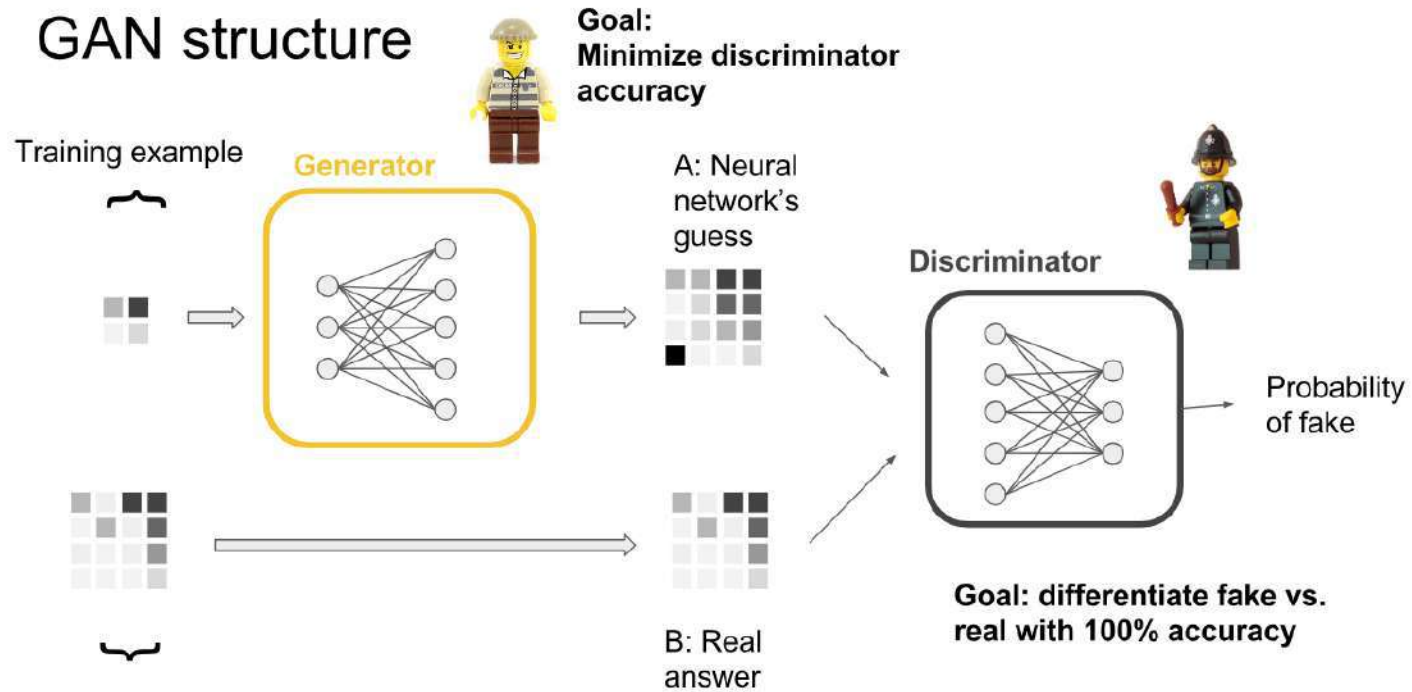


# China Social Score

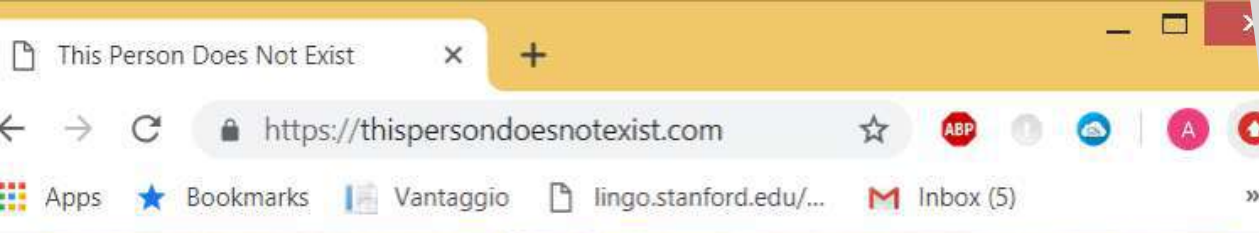
- Source:  
<https://www.wired.co.uk/article/china-social-credit-system-explained>



# Adversarial attack







Produced by a GAN  
StyleGAN (Der  
Original GAN  
Don't pan  
Help  
Ch

# Adversarial attack

<https://thispersondoesnotexist.com/>



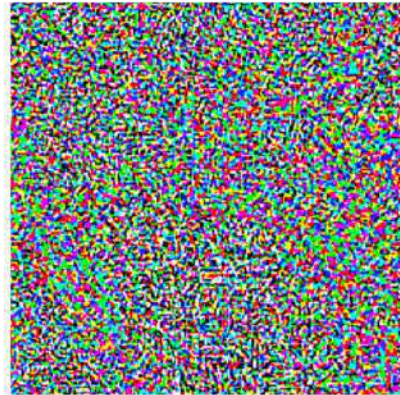
# Adversarial attack



“panda”

57.7% confidence

+ .007 ×



noise

=



“gibbon”

99.3% confidence

# Adversarial attack

---

(a) Image



(b) Prediction



(c) Adversarial Example



(d) Prediction





# Some applications

---



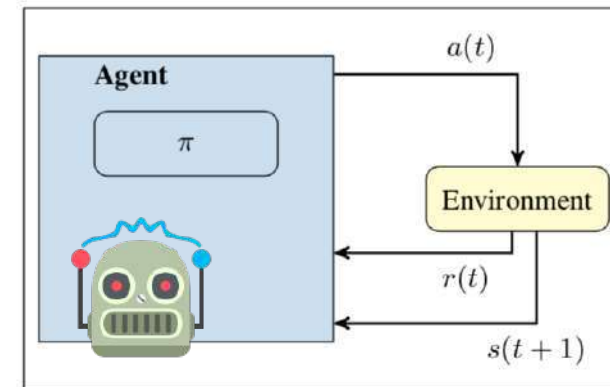
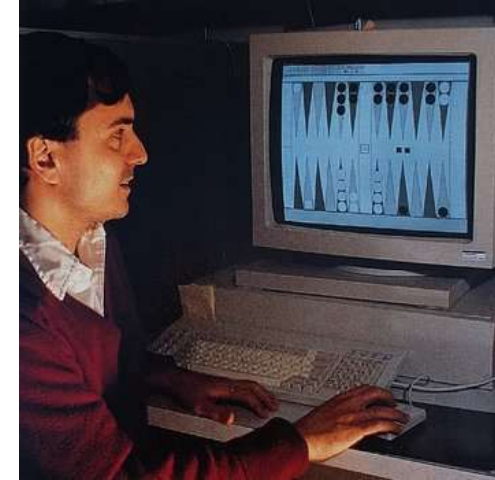
# Applications

- A Notion of Distance Between CP-nets
- Metric Learning for Value Alignment
- When is it morally acceptable to break the rule?
- Genetic Approach to the Ethical Knob



# Deciding and Learning

- AI systems increasingly make decisions that affect our lives (e.g. 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: get the most points, drive the best route...
  - Intend for actions to model the values of those deploying them.
- ***Ethically Bounded AI:*** understand and model human preferences and objectives; subsequently use these to control the actions and behaviors of autonomous agents.
- ***We model preferences and ethical priorities as CP-nets and propose novel machine learning techniques to judge decisions.***



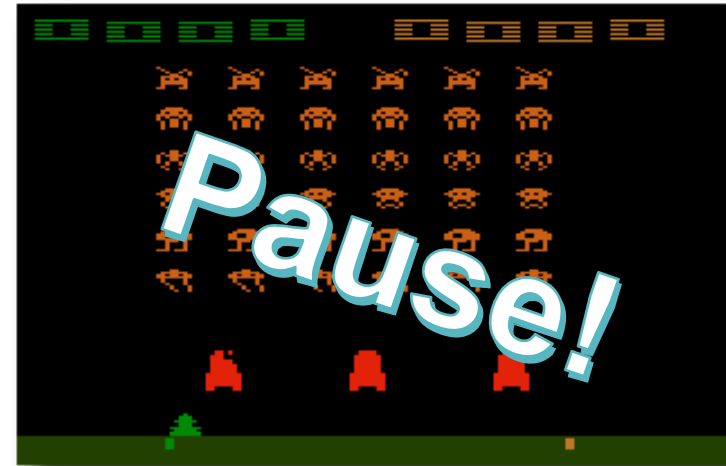
## Paper Citations

Francesca Rossi and Nicholas Mattei. *Building Ethically Bounded AI*, AAI 2019.

Francesca Rossi and Andrea Loreggia. 2019. Preferences and Ethical Priorities: Thinking Fast and Slow in AI. AAMAS 2019

# “Reward Hacking”

- Agents may “Reward Hack,” i.e., learn behaviors that have high reward but are not intended.
  - Constantly hitting the power-up instead of playing the game.
  - Pause the game instead of playing the game.
- One of a list of concrete problems in AI Safety including **Safe Exploration** and **Avoiding Negative Side Effects**.
- Wired Article: <https://www.wired.com/story/when-bots-teach-themselves-to-cheat/>
- DeepMind List: <https://t.co/mAGUf3quFQ>



**WIRED**

TOM SIMONITE BUSINESS 08.08.18 09:00 AM

**WHEN BOTS TEACH  
THEMSELVES TO CHEAT**

## Paper Citations

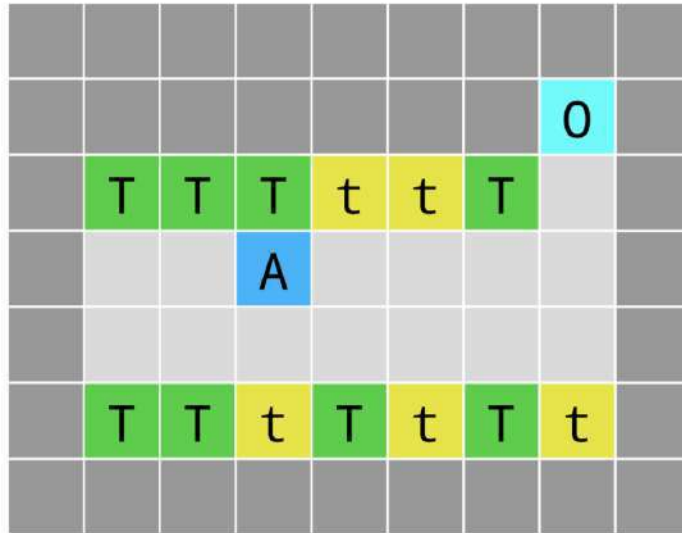
Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, Dan Mané.  
*Concrete Problems in AI Safety*. arXiv:1606.06565, 2016.

## Example

- Reinforcement learning agent goes in a circle hitting the same targets instead of finishing the race.
- <https://www.youtube.com/watch?v=tI0IHko8ySg&t=1s>



# Not Just Videogames!



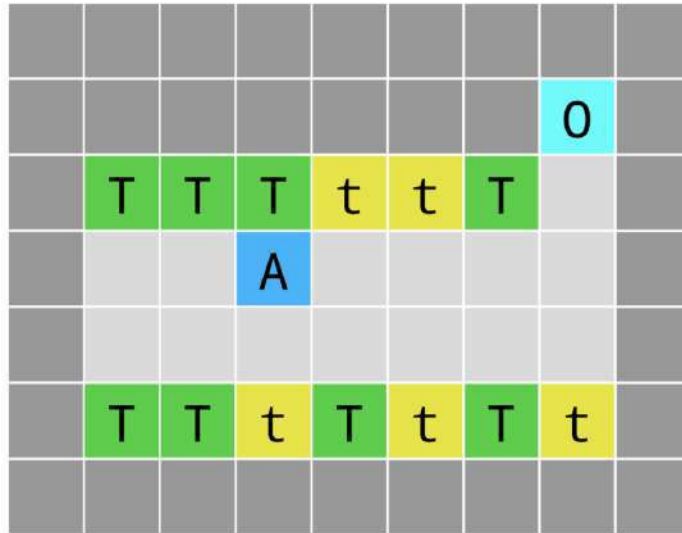
A	Agent
O	Bucket
T	Watered Tomato
t	Unwatered Tomato



- DeepMind and others released AI Safety Grid World posing a number of challenging RL tasks.
  - <https://arxiv.org/abs/1711.09883>
- Here we have a robot who must water the plants and is penalized if he sees a plant that is un-watered.



# Not Just Videogames!



A	Agent
O	Bucket
T	Watered Tomato
t	Unwatered Tomato

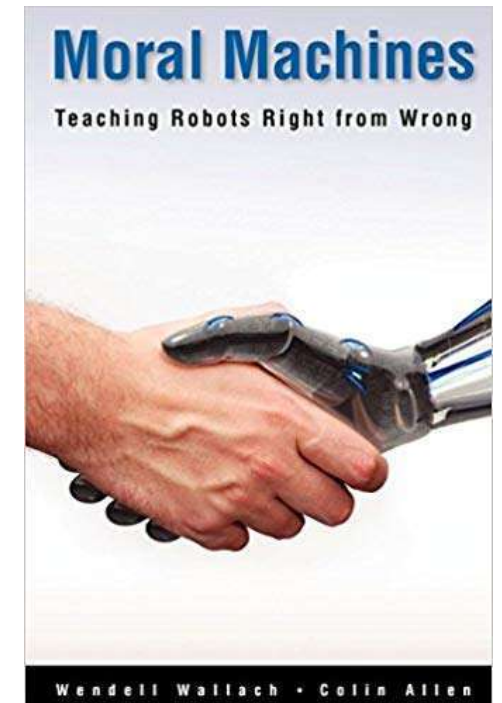


- DeepMind and others released AI Safety Grid World posing a number of challenging RL tasks.
  - <https://arxiv.org/abs/1711.09883>
- Here we have a robot who must water the plants and is penalized if he sees a plant that is un-watered.

# Ethically Bounded AI: Value Alignment and Machine Ethics



- In many settings we want to combine the creativity of AI with constraints that come from many places including ethics, morals, business process, guidelines, laws, etc.
- **Ethics v. Morality:** *mores or morals* are the customs, norms, or conventions of a particular community or society and *ethics* is a thoughtful, coherent reflection on, and application of, these norms [Michael J. Quinn, *Ethics for the Information Age*, 2015].
- Two main approaches:
  - **Top Down:** write down all the rules and have the agent follow them.
  - **Bottom Up:** show the agent appropriate actions.
- Key question: **How do we control the behavior of autonomous agents, without explicitly telling them what to do, so they comply with our constraints?**



## Paper Citations

Emanuelle Burton, Judy Goldsmith, Nicholas Mattei.

*How To Teach Computer Ethics with Science Fiction*. Communications of the ACM (CACM), 2018.

# Preferences in CS

- Preferences are a fundamental primitive that use to understand the intentions and desires of users.
  - Likes, stars, rankings, ratings.
- We also get detailed information from agents, systems, and algorithms that rank, sort, score, and combine judgments about actions and outcomes.

## {PrefLib}: A Library for Preferences

[Main](#)
[About](#)
[Papers](#)
[Data Formats](#)
[Data By Domain](#)
[Data By Type](#)
[Tools](#)

A reference library of preference data and links assembled by [Nicholas Mattei](#) and [Toby Walsh](#). We currently house over 3,000 datasets for use by the community.

We want to provide a comprehensive resource for the multiple research communities that deal with preferences, including computational social choice, recommender systems, data mining, machine learning, and combinatorial optimization, to name just a few.

Please see the [about](#) page for information about the site, contacting us, and our citation policy. We rely on the support of the community in order to grow the usefulness of this site. To contribute, please contact [Nicholas Mattei](#) at: [nicholas\(dot\)mattei@nicta.com.au](mailto:nicholas(dot)mattei@nicta.com.au)

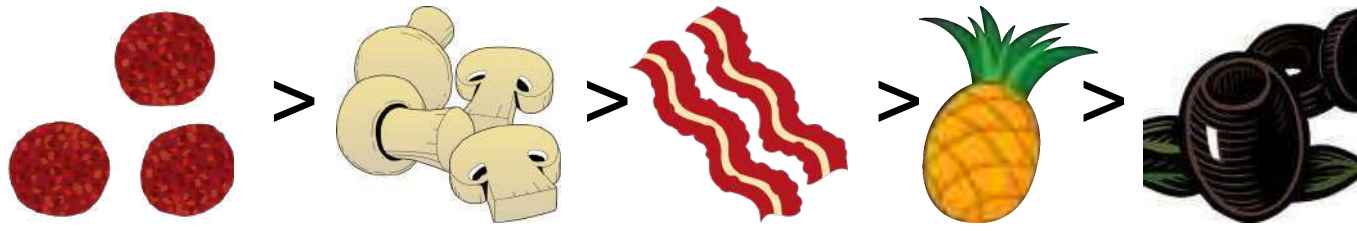
$a > b > c > d$

$\frac{1}{2} : a > b > e$   
 $\frac{1}{4} : c > b > a$   
 $\frac{1}{4} : b > c > a$

$a > b, c, d > e$

**Supported By:**

NICTA



**Sept. 3, 2013:**  
A big update today brings us over 3000 datasets hosted on the site with a full data archive over 7 GB!

We have also added a [Thanks!](#) section to recognize those individuals who have helped make PrefLib possible.

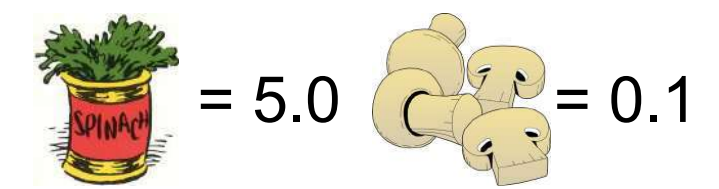
---

**July 1, 2013:**  
Our paper has been accepted to [2013 Conference on Algorithmic Decision Theory](#). We have also had several new donated datasets which have been parsed and posted.

We have added a new [Papers](#) section to the site with a list of papers that have used PrefLib!

**Links**

- [UC Irvine Machine Learning Repository](#)
- [University of Minnesota GroupLens Data Sets](#)
- [CSPLib: A Problem Library for Constraints](#)
- [Microsoft Learning to Rank Datasets](#)
- [SATLib: The Satisfiability Library](#)
- [Preference Learning.org](#)
- [Toshihiro Kamishima's Sushi Preference Dataset](#)
- [MAX-SAT Evaluations and Datasets](#)

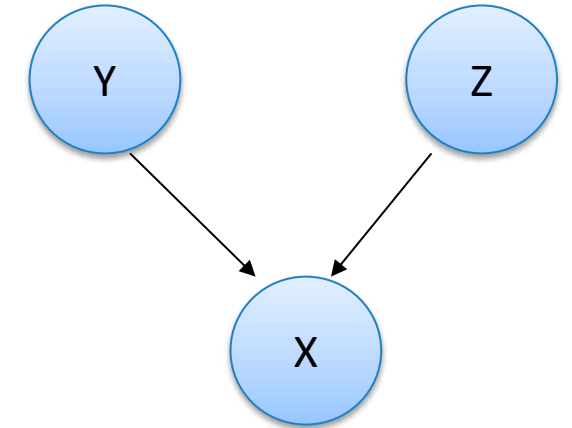


### Paper Citations

Nicholas Mattei and Toby Walsh.  
*PrefLib.Org: A Library for Preferences*. Proc. Algorithmic Decision Theory (ADT), 2013.  
*A PrefLib.org Retrospective: Lessons Learned and New Directions*. Trends in Computational Social Choice, Chapter 15, 2017.

# CP-Nets

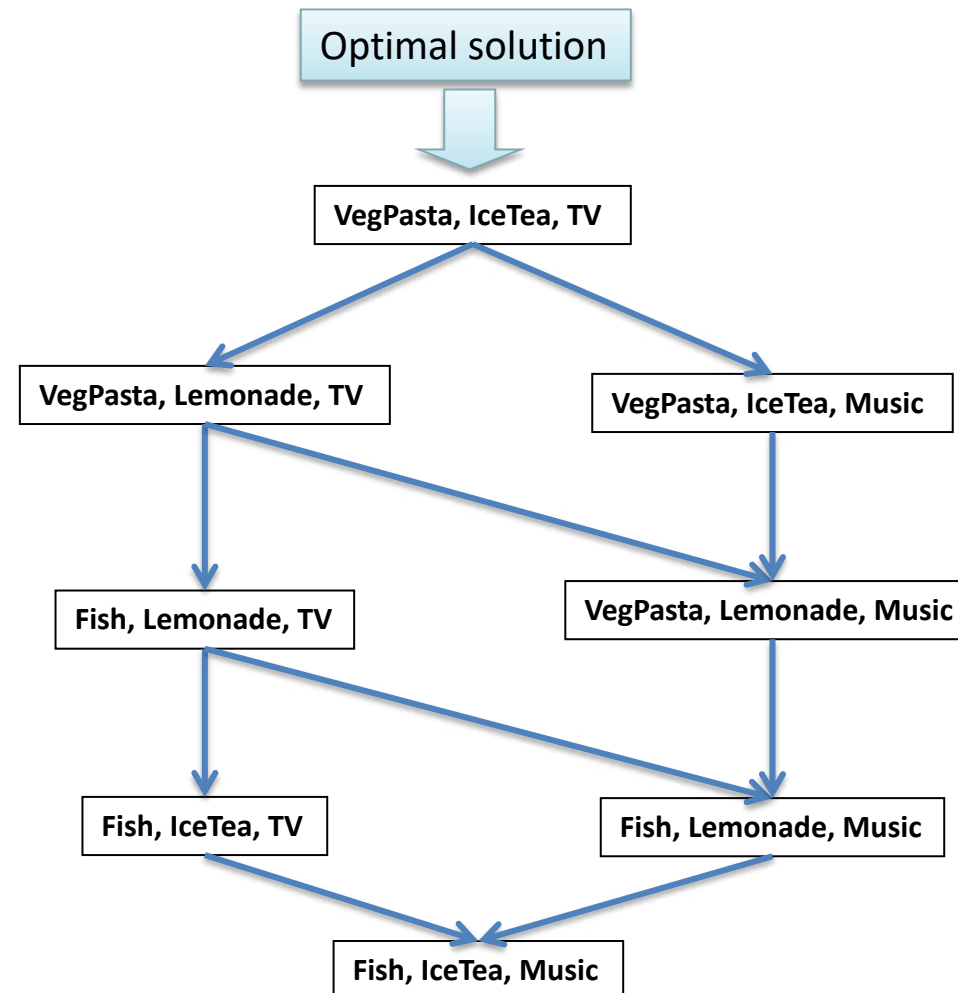
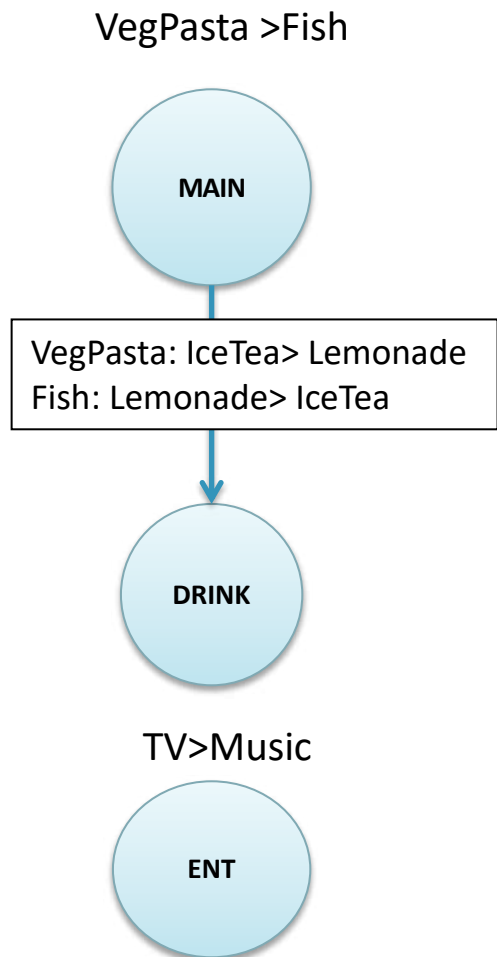
- Encode a subset of partial orders and follow the semantics of *all else being equal I prefer X to Y*.
- Variables  $\{X_1, \dots, X_n\}$  each with a possibly different domain.
- For each variable, a total order over its values
- **Independent variable:** a variable with no conditions.
  - $X := v_1 > v_2 > \dots > v_k$
- **Conditioned variable:** a total order for each combination of values of some other variables (conditional preference table)
  - $Y=a, Z=b: X=v_1 > v_2 > \dots > v_k$
  - X depends on Y and Z (parents of X)
- Graphically: **directed graph** over  $X_1, \dots, X_n$





# Example

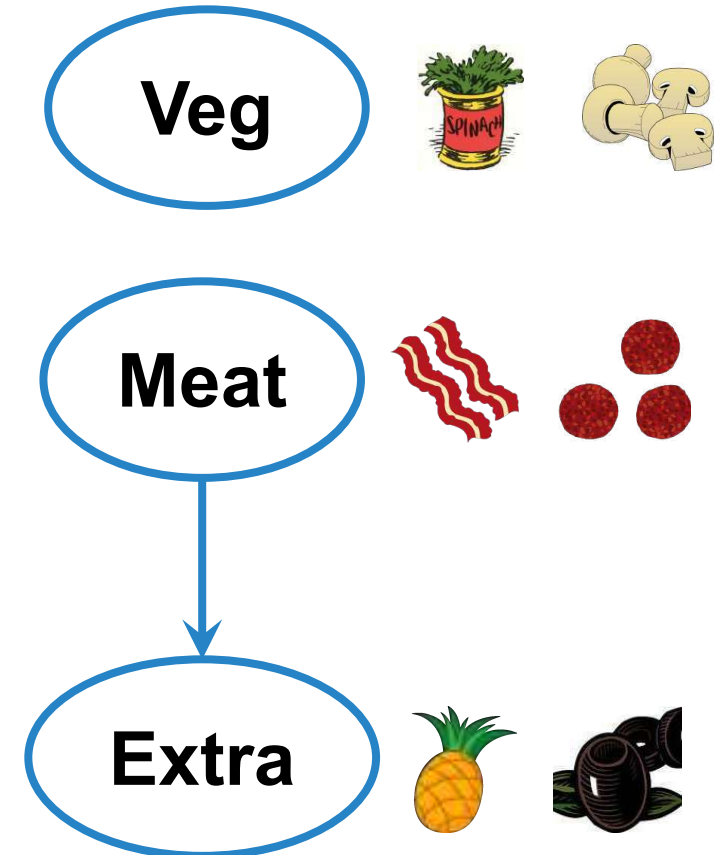
CP-net



Induced Ordering

# Distance Between Discrete Structures

- Preferences can take many forms: binary, scores, stars, orderings.
- Distances used in recommender systems (similarity of users), classification (distance to classes), and other places.
- **Distance (Metric):**
  - $d(x,y) \geq 0$  (non-negative),
  - $d(x,y) = 0$  iff  $x=y$  (identity),
  - $d(x,y) = d(y,x)$  (symmetry), and
  - $d(x,z) \leq d(x,y) + d(y,z)$  (triangle inequality).



## Paper Citations

Andrea Loreggia, Nicholas Mattei, Francesca Rossi, Kristen Brent Venable.

*On the Distance Between CP-nets*. Proc. Aut. Agents and Multiagent Systems (AAMAS) 2018.

*Value Alignment via Tractable Preference Distance*. Artificial Intelligence Safety and Security, Chapter 18, CRC Press, 2018.

*Preferences and Ethical Principles in Decision Making*. Proc. ACM/AAAI Conference on AI, Ethics, and Society (AIES), 2018.

*CPMetric: Deep Siamese Networks for Learning Distances Between Structured Preferences*. arXiv:1809.08350, 2018.

# Distance on partial orders

- Measure how similar/different are partial orders:
  - notion of distance over partial orders
- **Kendall's  $\tau$  with penalty parameter  $p$  (KT)**
  - Extends Kendall's  $\tau$  distance to partial orders
- Given two partial orders  $P$  and  $Q$  and two outcomes  $i$  and  $j$

$$KT(P, Q) = \sum_{i, j, i \neq j} K_{i, j}^p(P, Q)$$

where

$$K_{i, j}^p(P, Q)$$

**1** if  $i$  and  $j$  are ordered in the opposite way

**0** if  $i$  and  $j$  are ordered in the same way or incomparable in both POs

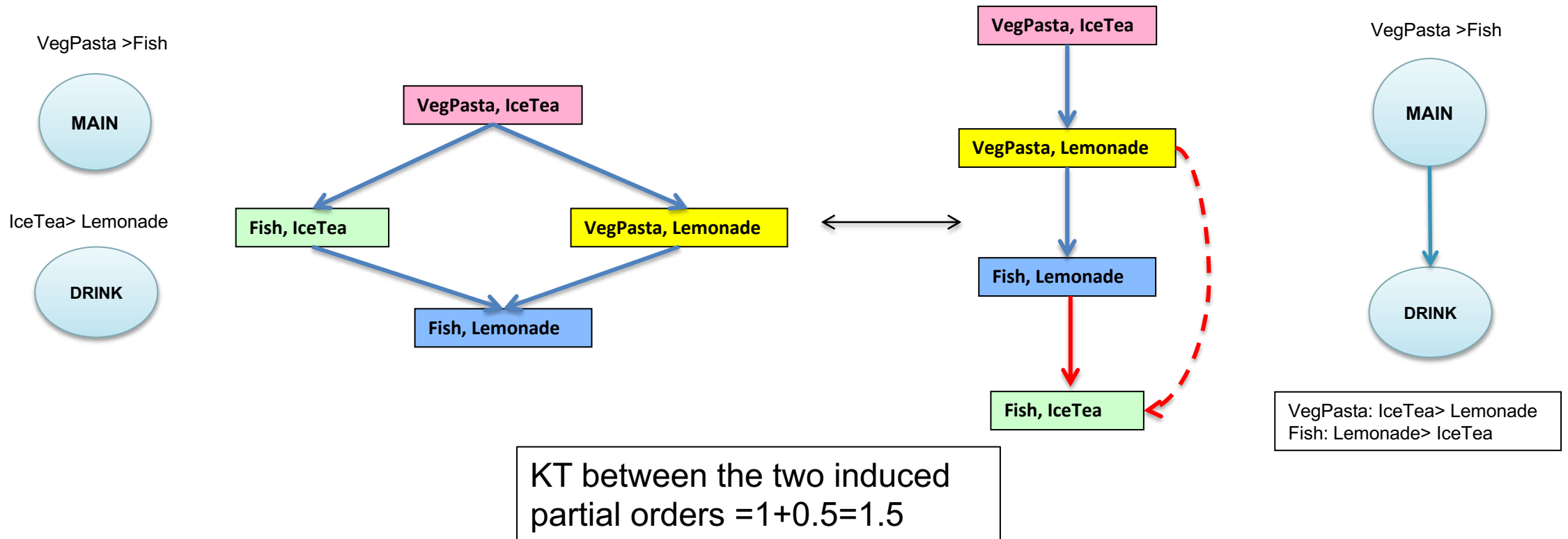
**$p$**  if  $i$  and  $j$  are ordered in one PO and incomparable in the other

## Paper Citations

Ronald Fagin, Ravi Kumar, Mohammad Mahdian, D. Sivakumar, and Erik Vee.  
*Comparing partial rankings*. SIAM J. Discret. Math., 20(3):628–648, March 2006.

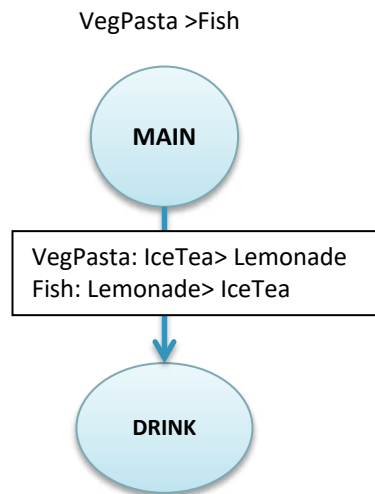
# Distance between Structures?

- Given two CP-nets defined over the same set of features, how similar/different are the preferences they represent?

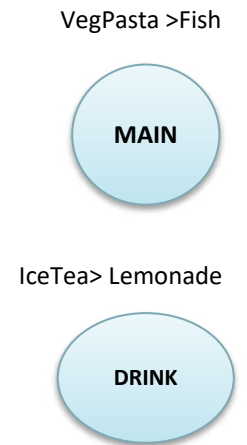


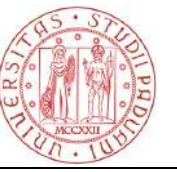


# Examples



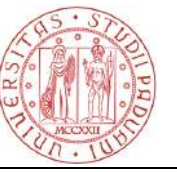
Can we compute the KTD distance directly from the CP-nets in polynomial time?





# Our setting

- The CP-nets we consider:
  - All the **same** set of **binary features**
  - **Acyclic**
  - **O-legal**: there is an ordering  $O$  of the features such that if there is an edge  $X \rightarrow Y$  in the CP-net, then  $X$  comes before  $Y$  in  $O$

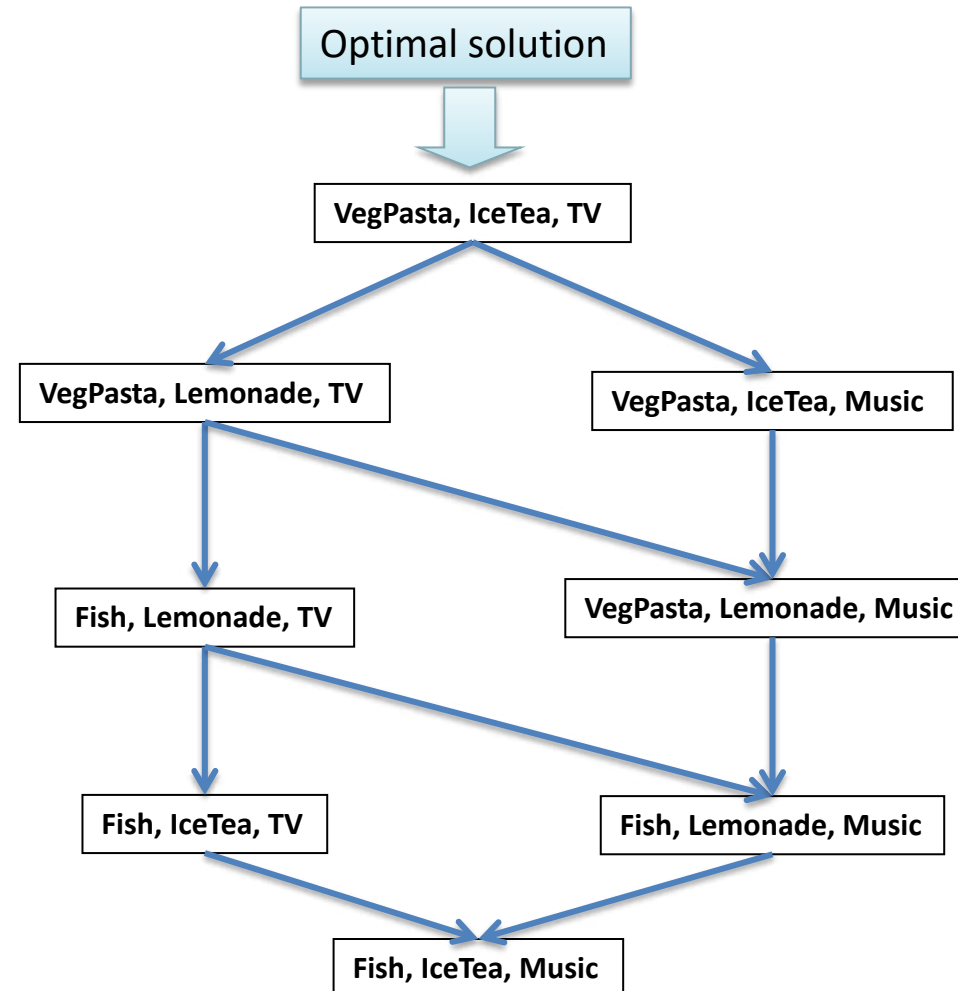
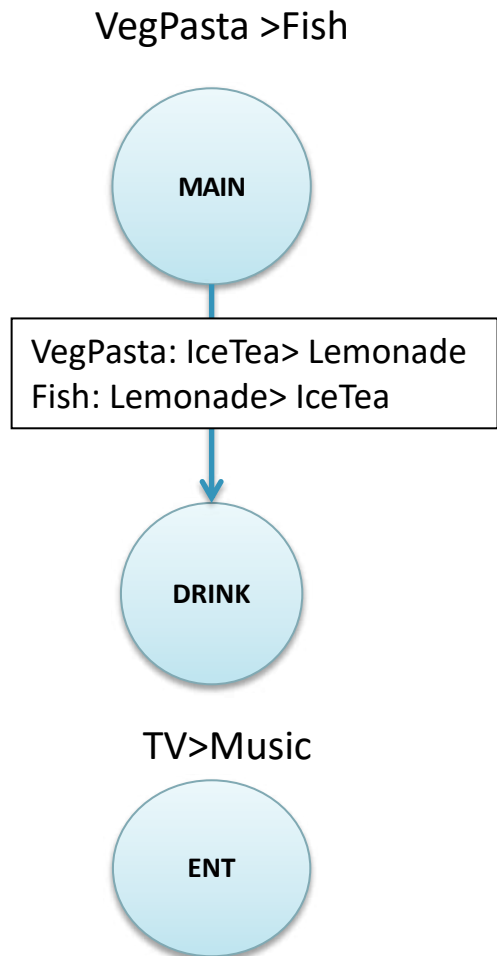


# Approximating the KTD distance

- Instead of computing the **KTD between two** CP-nets in polynomial time,
- Compute the **KT of two particular linearization of the POs** from the CP-nets in polynomial time
  - That is, without explicitly computing the linearizations!

# Example

CP-net

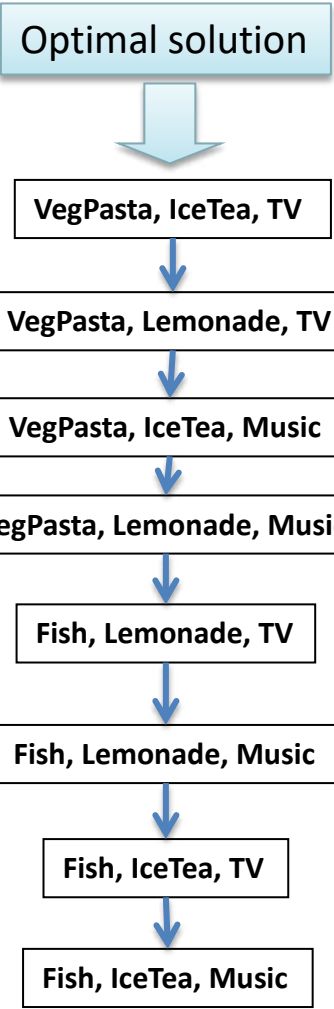
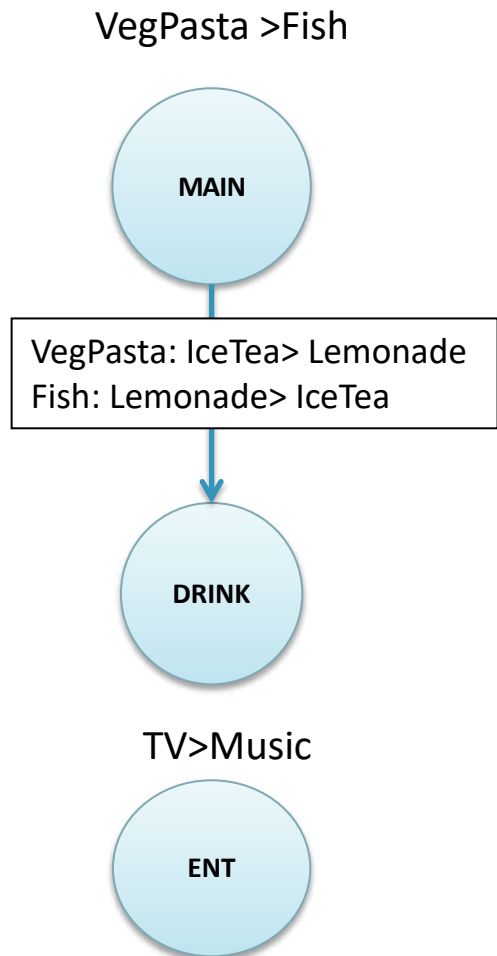


Induced Ordering



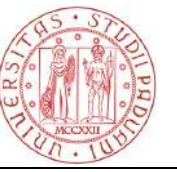
# Example

CP-net



Linearization

There are linearizations such that finding the Next best solution directly from the CP-net is easy (polynomial delay)



# CPD distance

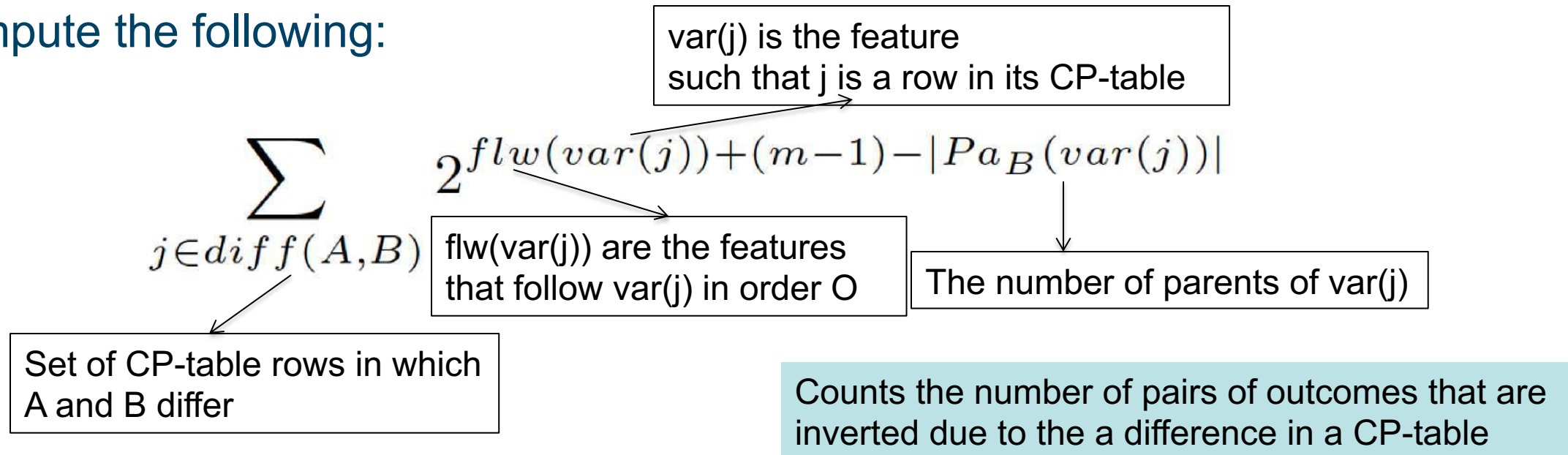
- Given two O-legal CP-nets A and B we denote with **LexO(A)** and **LexO(B)** the linearizations of their induced partial orders
  - as defined in Boutilier et al. 2004.
- We define:

$$\text{CPD}(A,B)=\text{KT}(\text{LexO}(A),\text{LexO}(B))$$

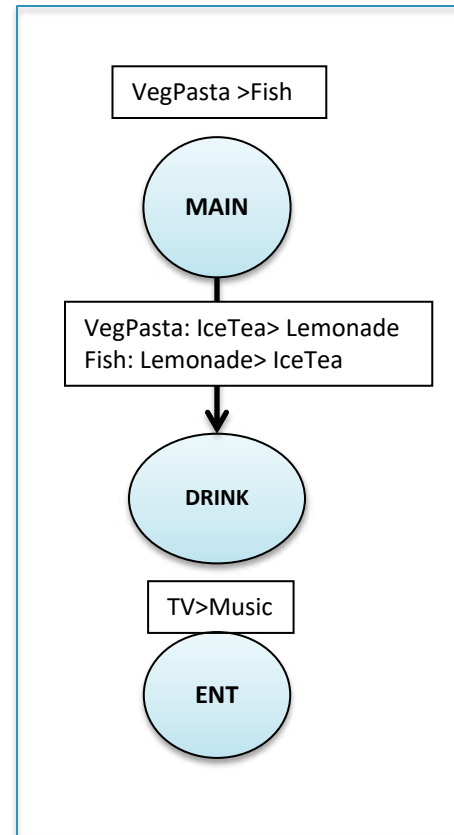
It is easy to see that CPD is a distance

# CPD: finding approximation

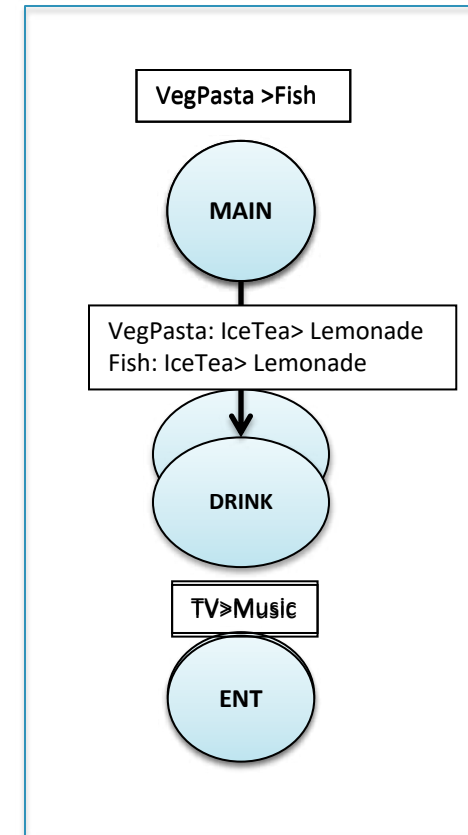
- Measuring the distance between CP-nets is exponential in the worst case.
- TH: **Given two O-legal CP-nets A and B, with  $m$  features,  $CPD(A,B)$  can be computed in polynomial time as follows:**
  1. **Normalize** A and B so that all features have as parents the union of their parents in A and B (redundant rows are added to the CP-tables)
  2. Compute the following:



# Computing CPD: Step 1 Normalization



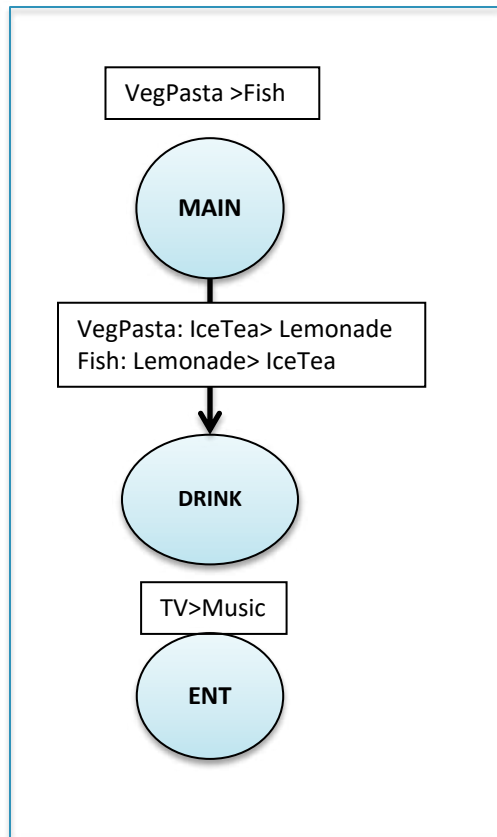
CP-net A



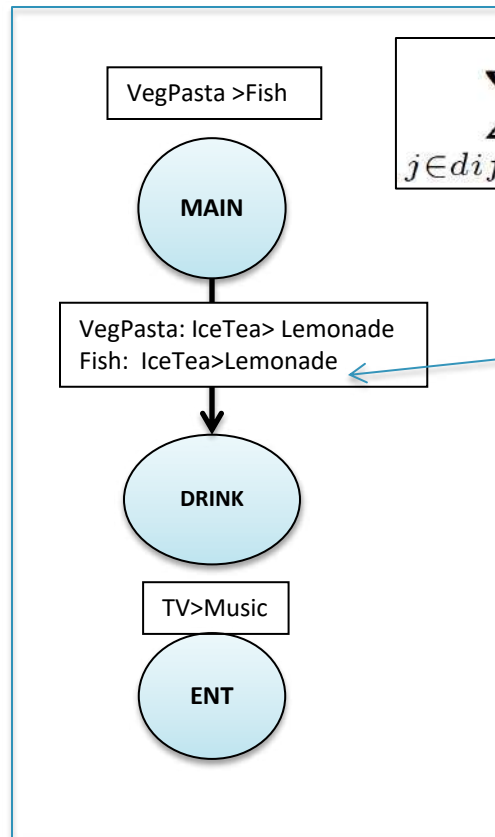
CP-net B



# Step 2: Count



CP-net A



CP-net B

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

diff(A,B)

var(j)=DRINK

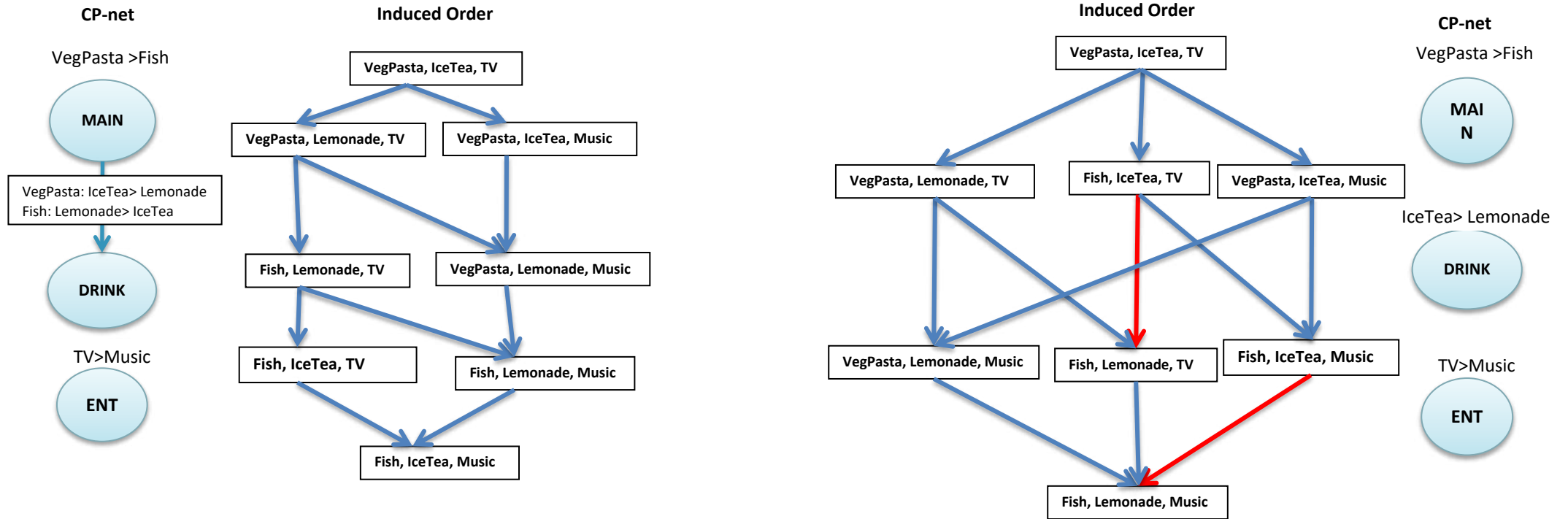
flw(DRINK)=1 (only ENT)

m=3

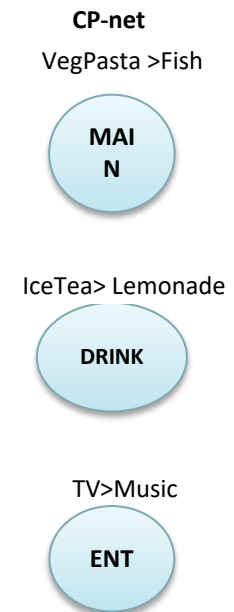
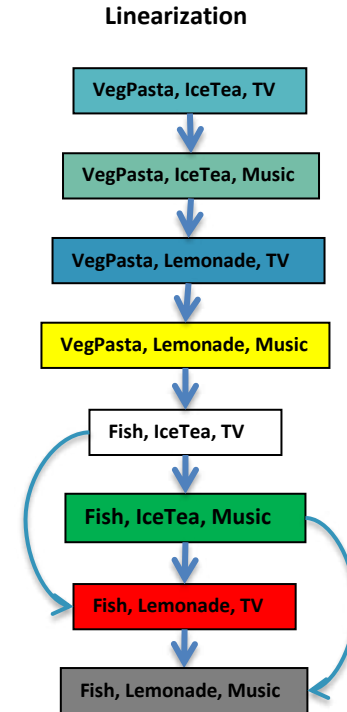
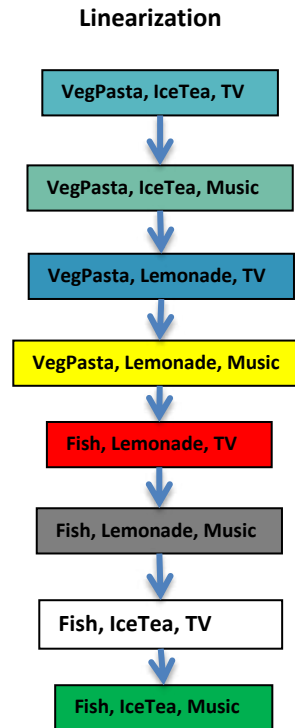
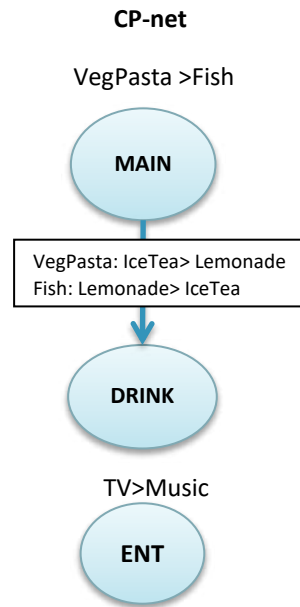
|PA(DRINK)|=1, DRINK has only MAIN as parent

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

# Examples



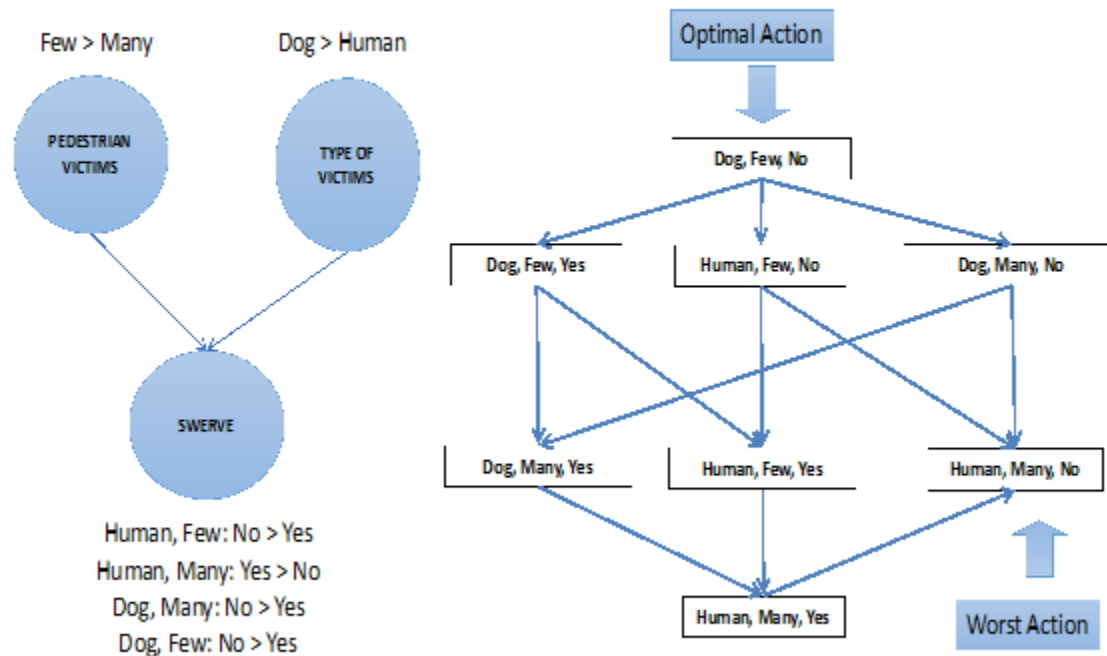
# Examples



# CP-nets as Ethical Priorities

- **Moral Preferences:** Amartya Sen, “morality requires judgment among preferences.”
  - Meta-ranking: preferences over preferences.
  - The preferences of an individual can be morally evaluated by measuring the distance of his/her CP-net from the moral one.

Collective Ethics



Angel Driver Ethics

Dog > Human



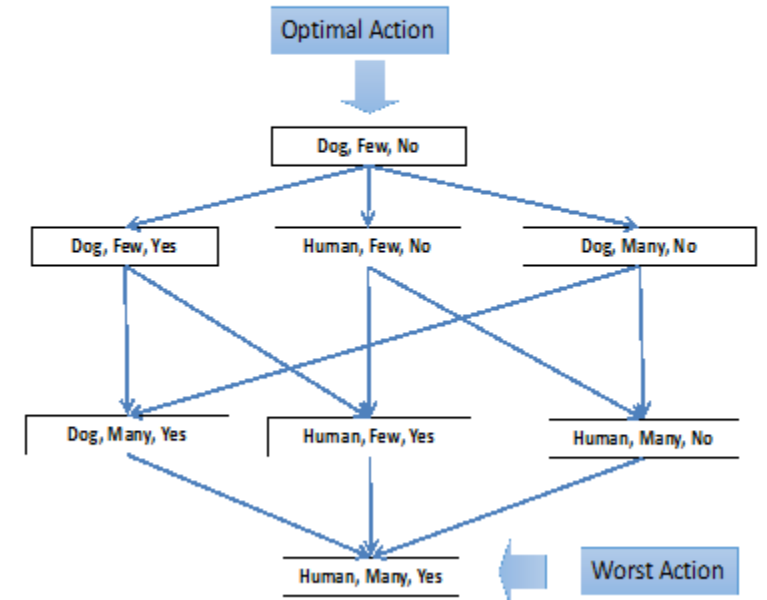
Few > Many



No > Yes



Induced Ordering

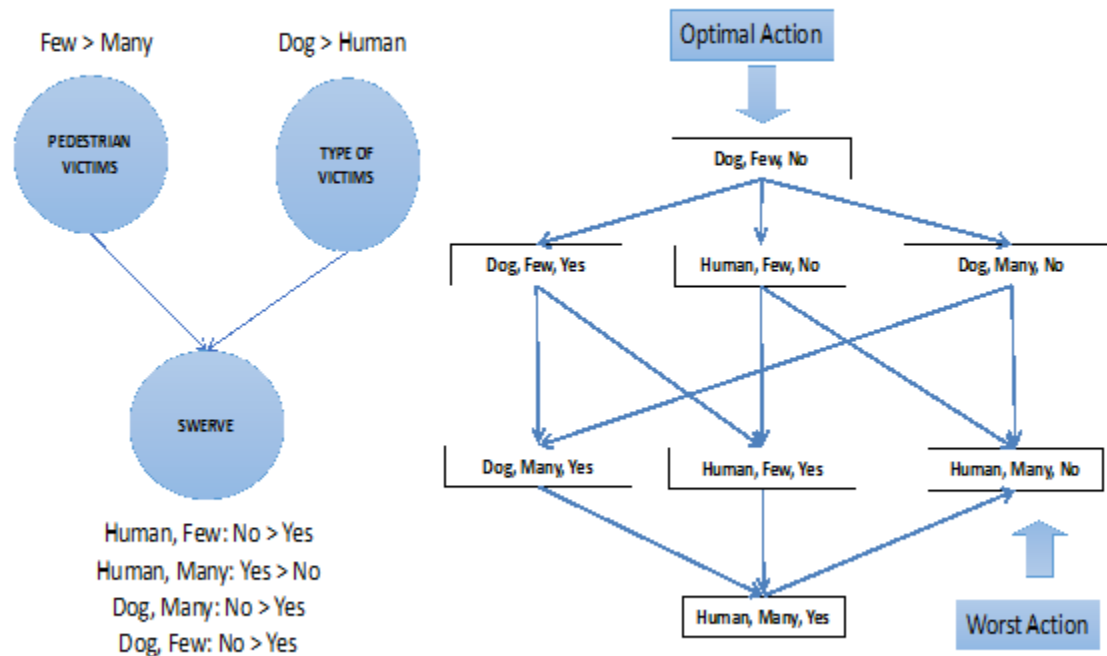




# CP-nets as Ethical Priorities

- **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.

## Collective Ethics



## Angel Driver Ethics

Dog > Human



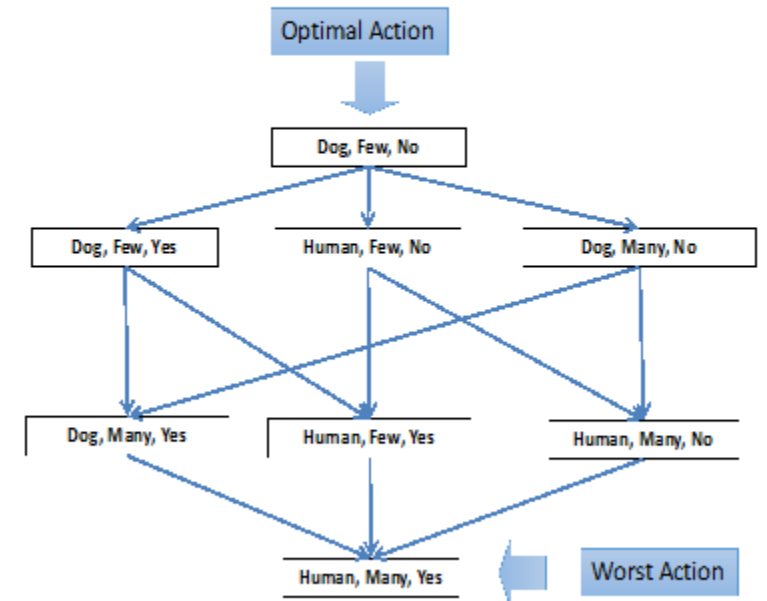
Few > Many



No > Yes



## Induced Ordering





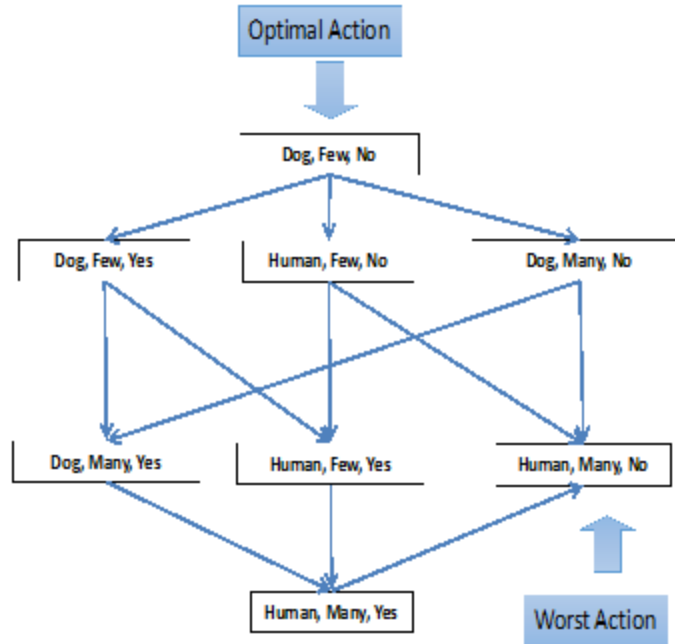
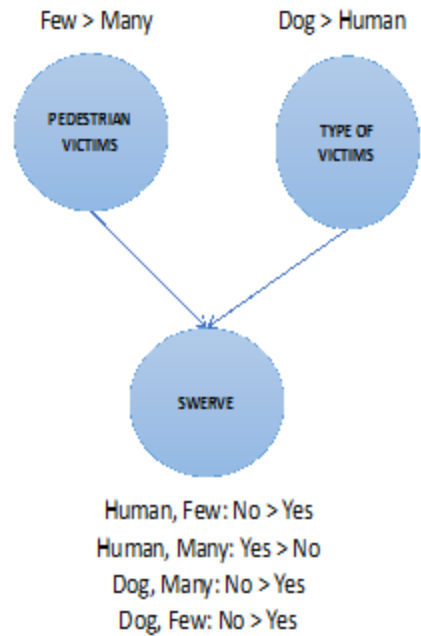
# Value Alignment Procedure

---

- Given an ethical principles and individual's preferences.
  - A. Set two distance thresholds:  $t_1$  (ranging between 0 and 1) between CP-nets, and  $t_2$  between decisions (ranging between 1 and  $n$ )
  - B. 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
  - C. If so, individual is allowed to choose the top outcome of his preference CP-net
  - D. If not, then 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.

# Value Alignment Procedure

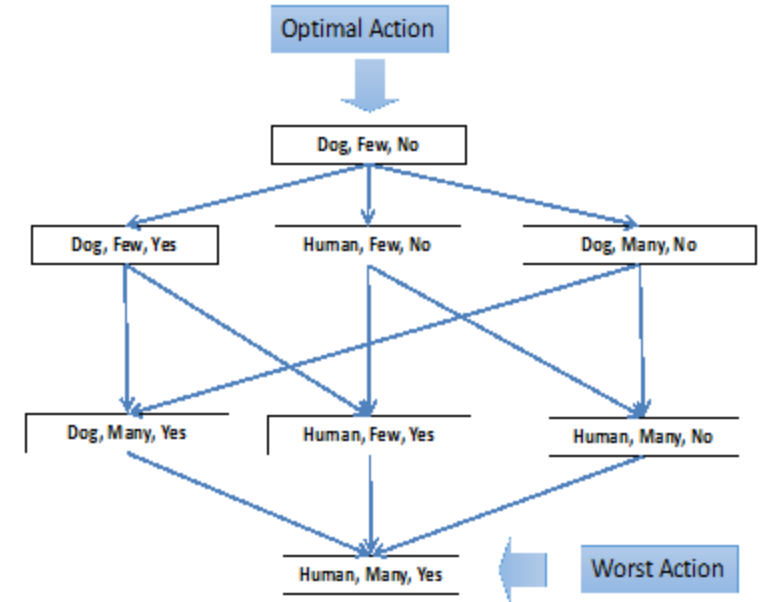
## Collective Ethics



## Angel Driver Ethics

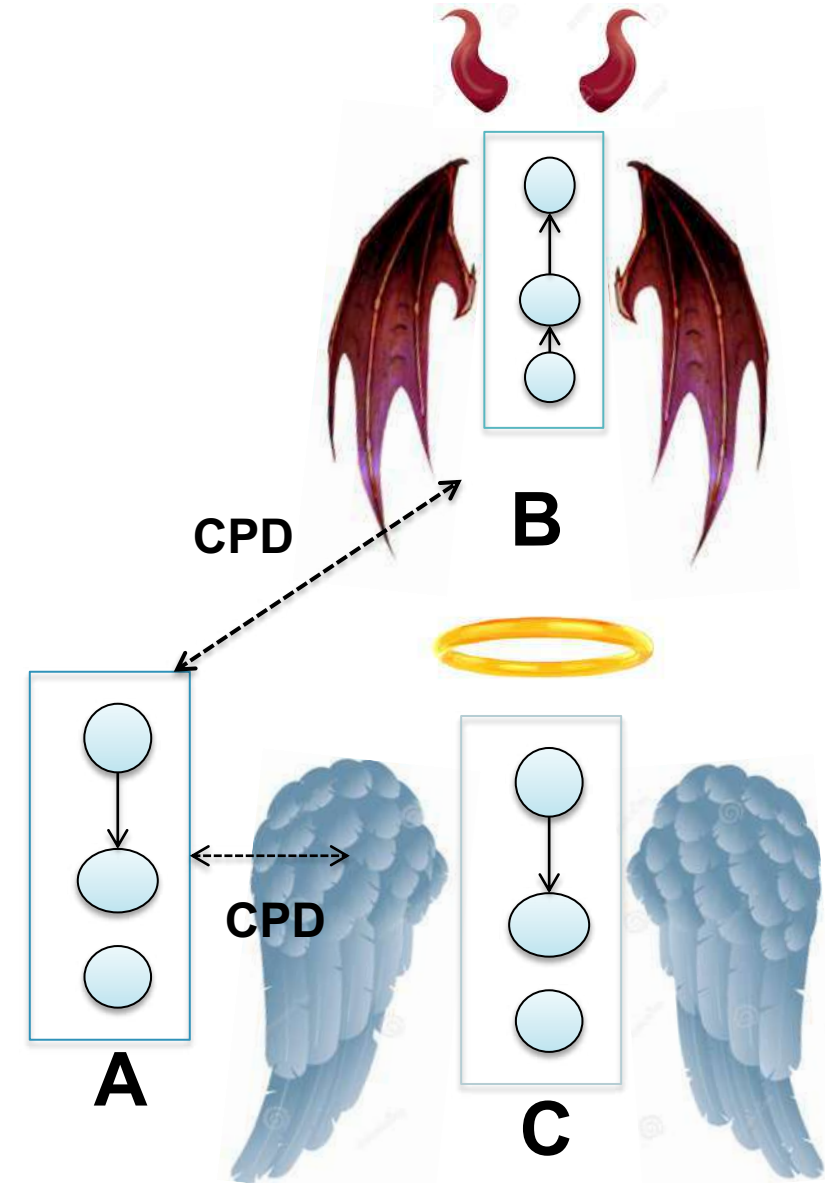
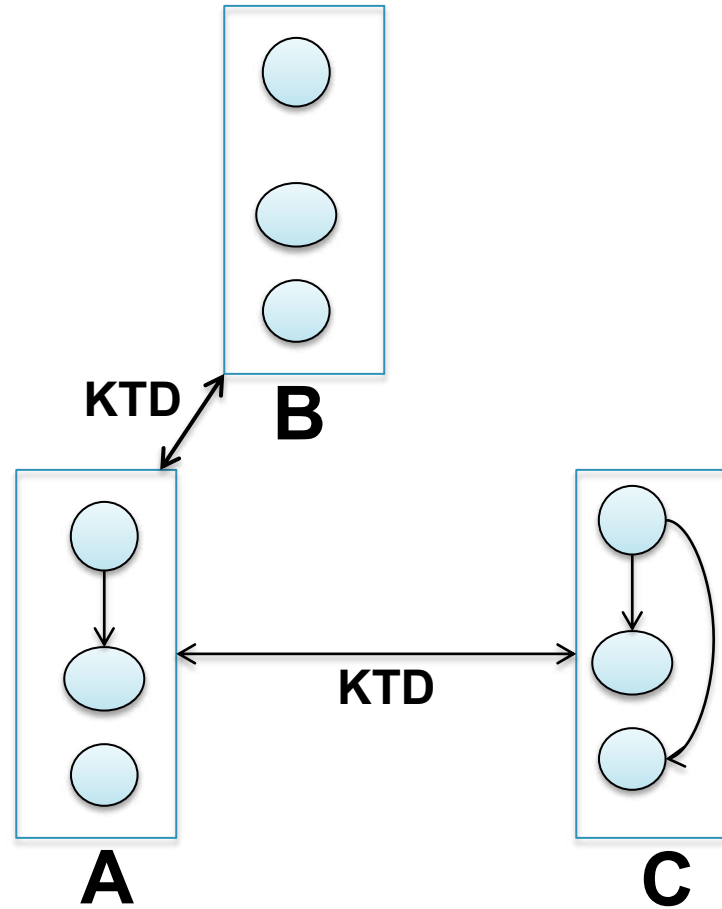


## Induced Ordering



# Value Alignment

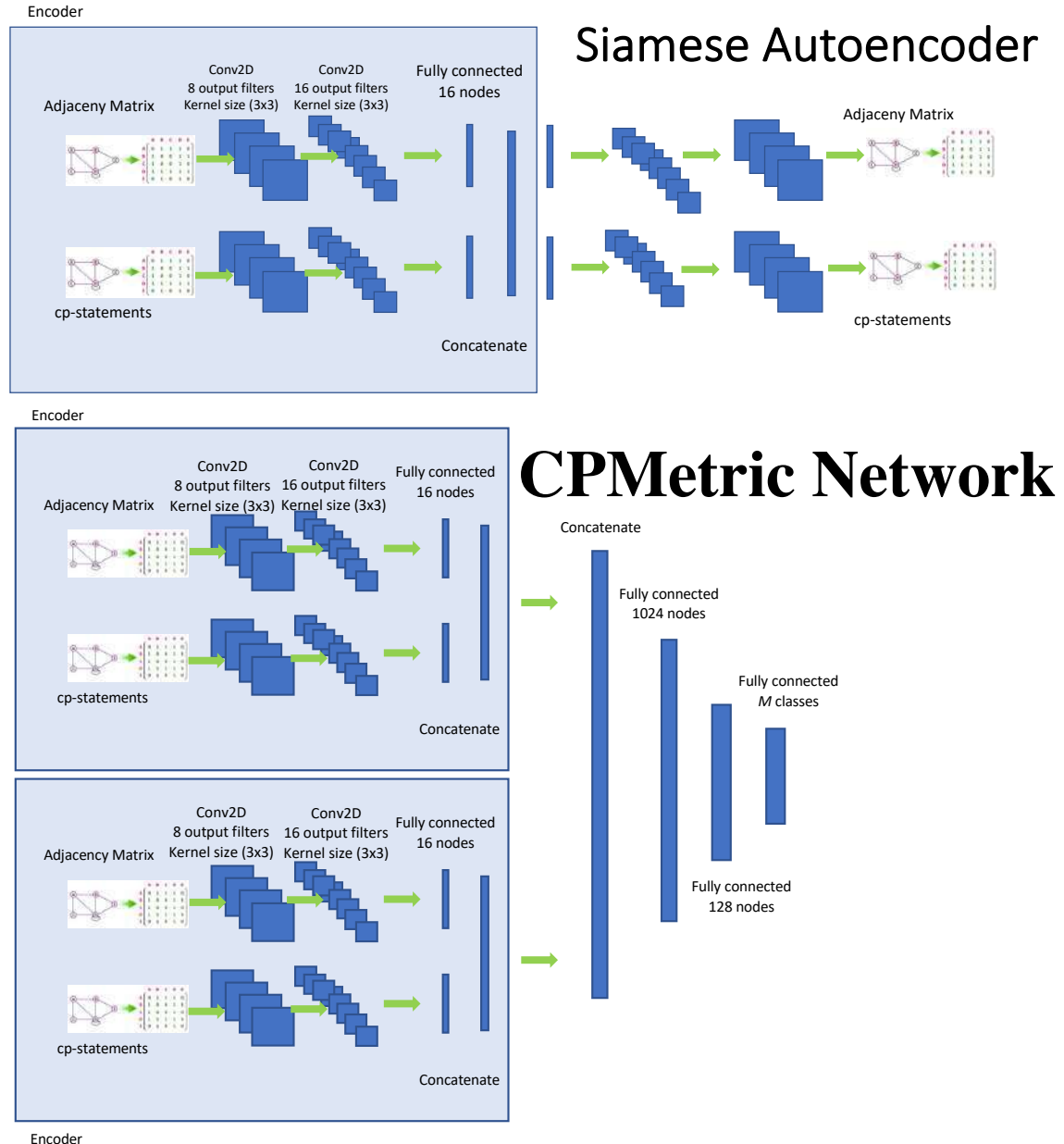
- We generate triplets of CP-nets (A,B,C).
- We chose one as pivot: A.
- We count how many time KTD says B is closer and the other distances say C is closer.
- CPD
- Gives us a notion of a “more compliant” CP-net.





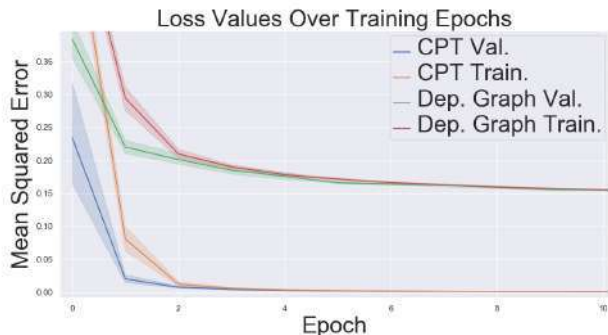
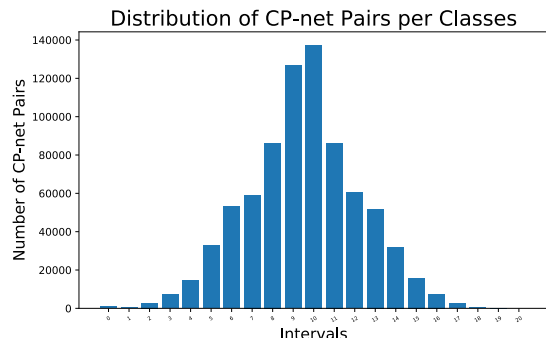
# Measuring the Distance

- Measuring the distance between CP-nets is exponential in the worst case.
- Need to find a way to evaluate the distance between, e.g., two competing CP-nets and a third “Moral” CP-net. Judge which one is “more aligned.”
- Using machine learning we have two steps:
  - Encode the CP-net (graph embedding issues).
  - Determine the distance.
- We encode the normalized laplacian matrix of the graph and a table of the cp-statements.



# Experiments and Results

- For training we generate 1000 randomly generated CP-nets and compute the distance for all pairs for all  $n = \{3, \dots, 7\}$ . For testing we generate another 1000 randomly generated CP-nets and find all possible triples.
- We get good convergence in the training phase and are able to learn a high quality latent representation.
- For the comparison task we are slightly outperformed by an approximation method, though we run two orders of magnitude faster.



	No Autoencoder	Autoencoder	Siam. Autoencoder	<i>I</i> -CPD
N	Accuracy on Triples	Accuracy on Triples	Accuracy on Triples	Accuracy on Triples
3	85.01% (2.01%)	85.76% (2.29%)	85.47% (2.32%)	91.80%
4	91.17% (0.92%)	91.38% (1.10%)	91.78% (1.13%)	92.90%
5	88.40% (0.91%)	89.36% (1.08%)	89.18% (1.08%)	90.80%
6	87.33% (0.80%)	87.17% (1.33%)	86.79% (1.84%)	90.10%
7	84.79% (1.16%)	84.57% (1.14%)	85.12% (0.86%)	89.90%

Table 1: Performance of the various network architectures on the qualitative comparison task as well as performance of *I*-CPD. While our networks do not achieve the best performance on this task they are competitive with the more costly approximation algorithm *I*-CPD.



# Conclusions and Next Steps

- *We model preferences and ethical priorities as CP-nets and propose novel machine learning techniques to judge decisions.*
- Important Questions and Next Steps:
  - How do we measure distance between heterogenous structures?
  - How do we capture and encode norms/values/expectations?
  - How do we account for edge effects?
  - How do we transition our techniques to other preference representations / formalisms?

IBM researchers train AI to follow code of ethics

BEN DIXSON, TECHTALKS @RENDERERS JULY 16, 2018 2:28 PM



Image Credit: Jyoti Williams RUS / Shutterstock

In recent years, artificial intelligence algorithms have become very good at recommending content to users — a bit too good, you might say. Tech companies use AI to optimize their recommendations based on how users react to content. This is good for the companies serving content, since it results in users spending more time on their applications and generating more revenue.

IBM explores the intersection of AI, ethics and Pac-Man

**FAST COMPANY**

CO-DESIGN | TECH | WORK LIFE | CREATIVITY | IMPACT | AUDIO | VIDEO | N


10.25.18 | INNOVATION ENGINE

## IBM explores the intersection of AI, ethics and Pac-Man

The lessons you learn teaching an iconic video game character to rack up points—without mowing down ghosts—might serve a higher purpose.

SCORE: 656 CONSTRAINT 1/08

[Animation: courtesy of IBM]



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

---

Edmond Awad, Sydney Levine, **Andrea Loreggia**, Nicholas Mattei, Iyad Rahwan, Francesca Rossi, Kartik Talamadupula, Joshua Tenenbaum and Max Kleiman-Weiner







# 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



# Deontology

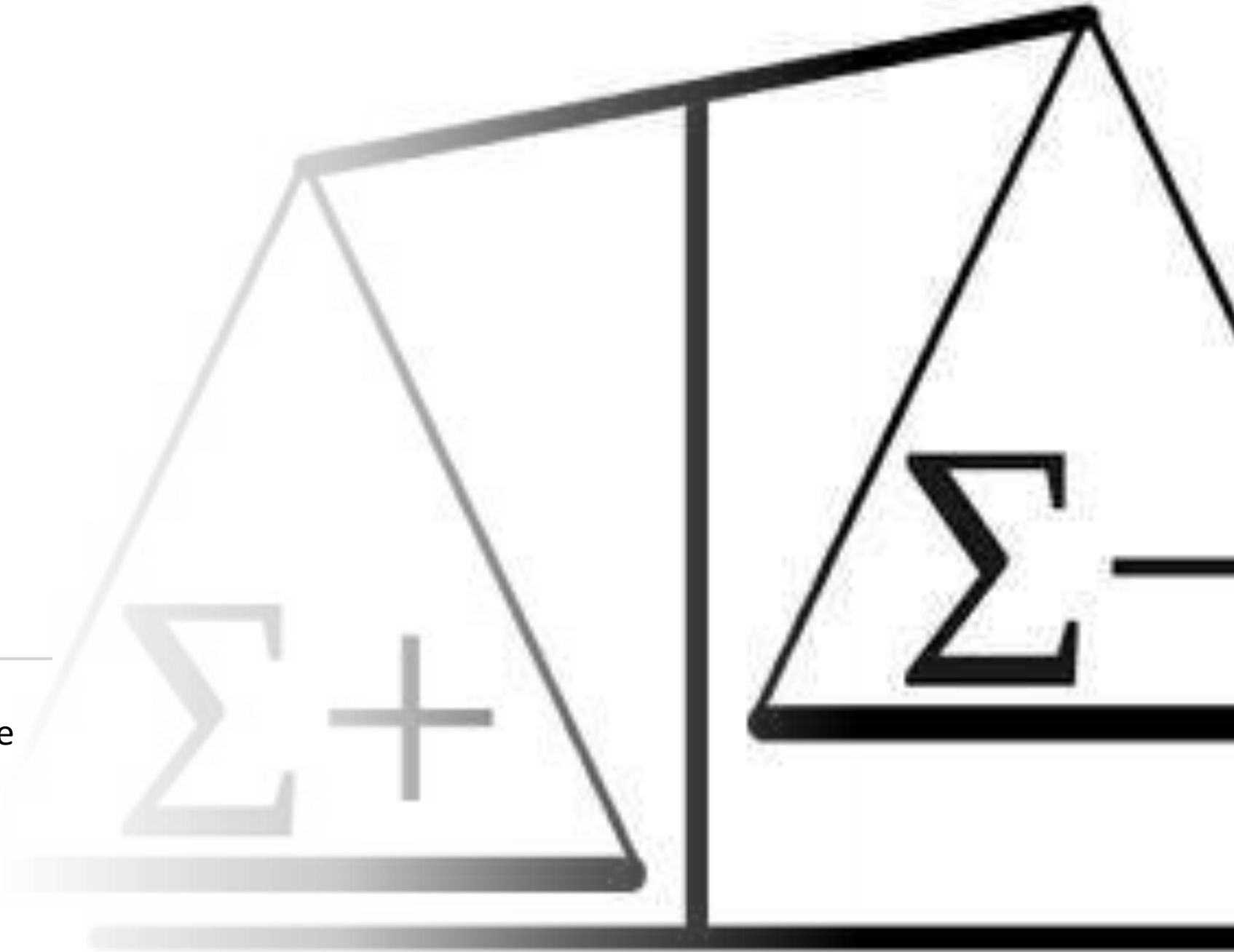
Following common rules that have been agreed upon by us or society



# Utilitarianism

---

Evaluating the consequences of the possible actions before deciding





# Contractualism

Finding an agreement between the parties involved









# Counter-examples

Under certain conditions, we are allowed to cut to the front of the line without waiting



# Triple Theory

A unified theory of moral cognition to:

- Combine elements of each of the theories of moral philosophy
- Build a computational model to direct actions of an AI system.





# Ethical Reasoning in AI Systems

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



# Experimental Details

- 27 short vignettes about people waiting in line in three different contexts (deli, bathroom, airport)
- 320 subjects were recruited from Amazon MTURK
- Subjects were randomly assigned to one of two experimental groups (moral judgment or context evaluation)

# Experimental Details

Moral judgment group:

- Read all the scenarios (27 total)
- For each scenario answer whether it was acceptable for the protagonist to cut in line (yes/no).

# Experimental Details

Context evaluation group:

- Subjects evaluated all the vignettes in one context only (9 questions).

# Experimental Details

Example of evaluation:

- Everyone: Think about the well-being of all the people in line combined. How are they affected by the person cutting in line?
- First Person: How much worse off/better off is the first person in line?



# Example

Imagine that there are five people who are waiting in line at a deli to order sandwiches for lunch. There is only one person (the cashier) working at the deli.

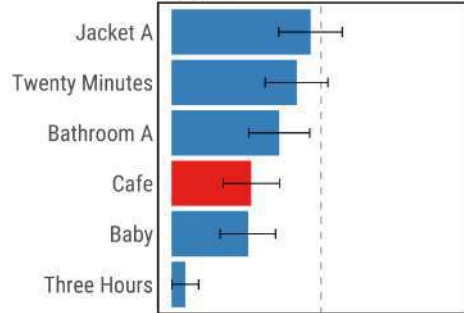
A customer who is eating lunch at the deli wants more a refill on tap water.

Is it OK for that person to ask the cashier for more water without waiting in line?

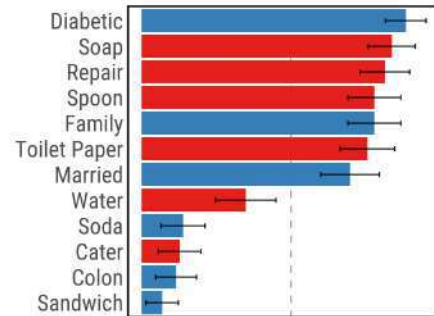
MENTIMETER

Requesting the main service? ■ No ■ Yes

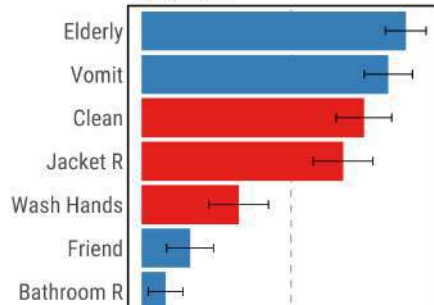
### Airport



### Deli



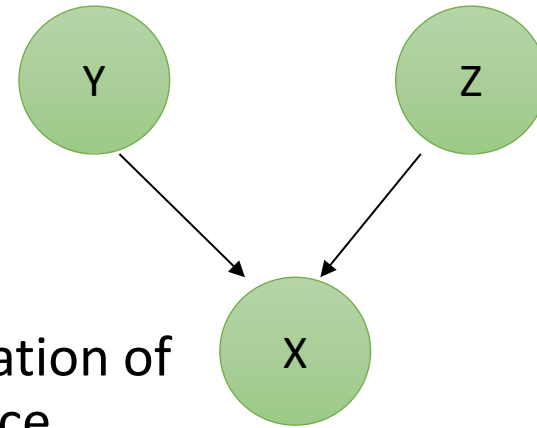
### Restroom



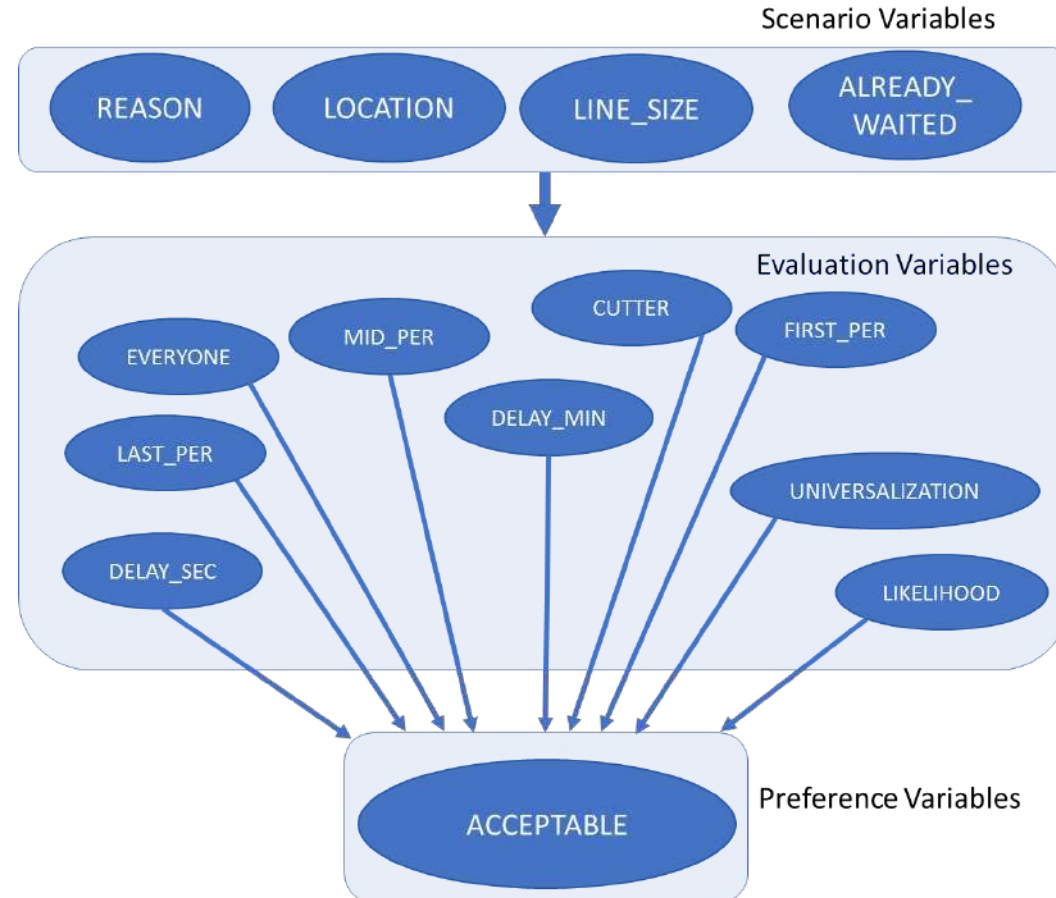
# Is it OK to cut the line?

# CP-nets

- Variables  $\{X_1, \dots, X_n\}$  with domains
- For each variable, a total order over its values
- **Independent variable:**
  - $X=v_1 > X=v_2 > \dots > X=v_k$
- **Conditioned variable:** a total order for each combination of values of some other variables (conditional preference table)
  - $Y=a, Z=b: X=v_1 > X=v_2 > \dots > X=v_k$
  - X depends on Y and Z (parents of X)
- Graphically: **directed graph** over  $X_1, \dots, X_n$ 
  - Possibly cyclic



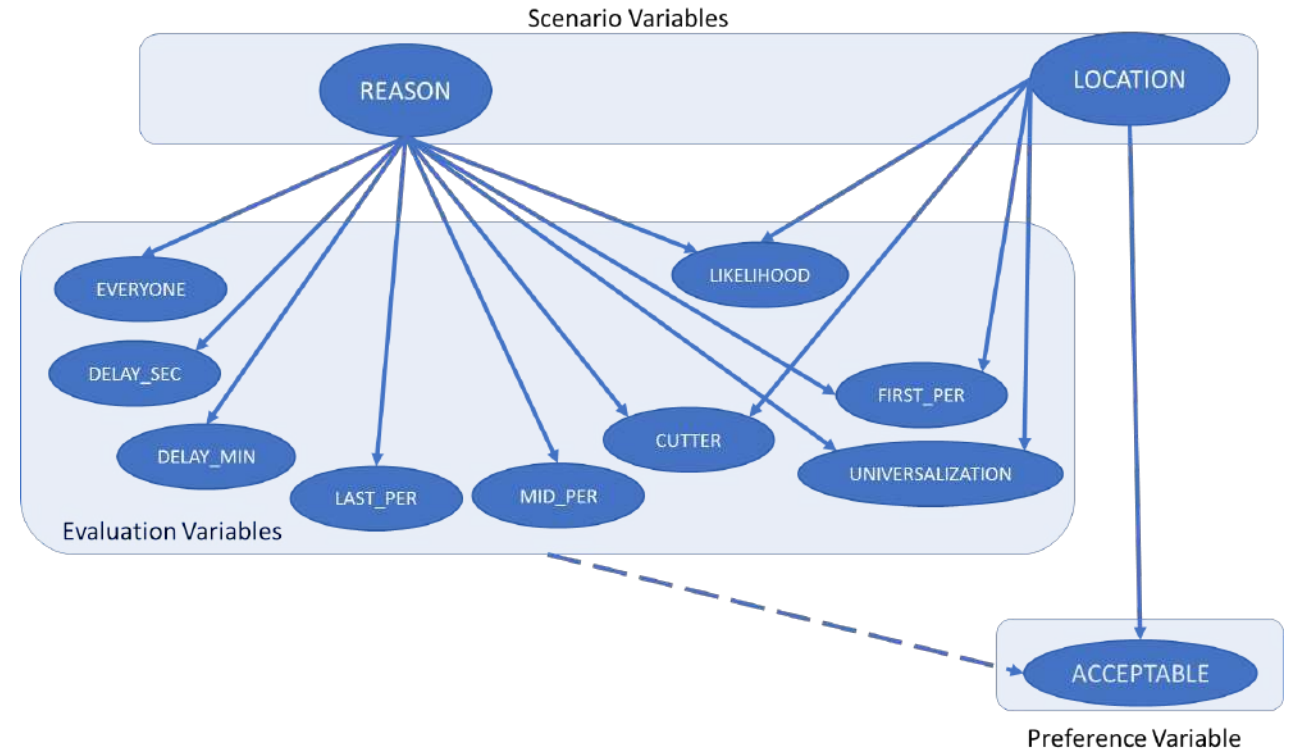
# Modelling and Reasoning with Preferences





# Data Analysis

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



# On-Going and Future Work

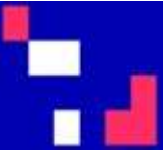
---

- Generalizing CP-nets to Model Moral Preferences
- Prescriptive Plans Based on Moral Preferences



# Conclusio

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



# Deontology/Kantian ethics

Giovanni Sartor





# Deontology

- Consequentialists hold that choices—acts and/or intentions—are to be morally assessed solely by the states of affairs they bring about.
  - E.g. my act of lying is good or bad depending on the effects it brings in the world
- Deontologists hold that certain actions are good or bad regardless of their consequences
  - Lying is always bad, regardless of its effect.
- The right has priority over the good: what makes a choice right is its conformity with a moral norm which orders or permits it, rather than its good or bad effect.
  - E.g. we should not kill anybody, even in those cases in which killing somebody would provide more utility. Is this always the case
    - Consider the case of the British soldier who apparently met Hitler in the trenches of 1<sup>st</sup> world war
    - What would a rule utilitarian say in such a case?
- The 10 commandments?

# Some ideas for being impartial

## Ethics and impartiality

- Is ethics linked to ideas of fairness or impartiality?
- Is it unethical to have a preference for oneself (or one's friends)?

## What about the golden rule

- Treat others as you would like others to treat you
- Do *not* treat others in ways that you would *not* like to be treated
- What you wish upon others, you wish upon yourself

## Is the golden rule useful

- Always? Can you find counterexamples?
- Would you want an AI system that applies it (with regard to its owner)?

# Immanuel Kant

- One of the greatest philosophers of all times
- Lived in Prussia (1724-1804)
- Addressed
  - The theory of knowledge: Critique of pure reason
  - The theory of morality: Critique of practical reasons
  - The theory of aesthetics (art): Critique of judgment
  - Law, logic, astronomy, etc.



# Kant's ethic and the principle of universalizability

- “Act only according to that maxim by which you can at the same time will that it should become a universal law” (1785).
- What is a maxim: a subjective principle of action, it connects an action to the reasons for the action (an intention to perform an action for a certain reason)
  - I shall donate to charities to reduce hunger
  - I shall deceive my contractual partner, to increase my gains
  - I shall cheat on taxes, to keep my money
  - I shall tell the truth, to provide trust
- Are they universalizable? Would I want them to become universal laws, that are applied by everybody?



# An universalisation test

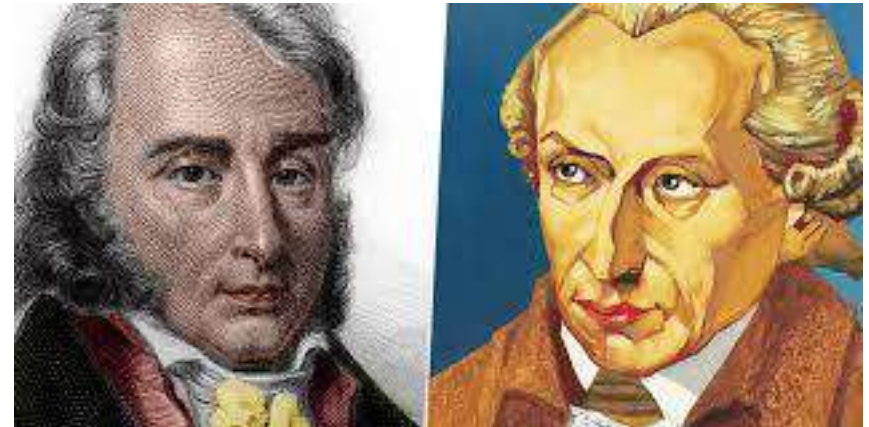
- Shafer Landau. The test of universalizability:
  - Formulate your maxim clearly state what you intend to do, and why you intend to do it.
  - Imagine a world in which everyone supports and acts on your maxim.
  - Then ask: Can the goal of my action be achieved in such a world?
- The process ensure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

Apply this principle to

- Cheating in an exam, in order go get a good mark
  - Giving money to a charity to relieve
- Would we want a robot following this maxim?

# Immanuel Kant vs Benjamin Constant

- Should one must (if asked) tell a known murderer the location of his prey.
  - It is ok to refuse to answer?
  - It is ok to tell a lie (e.g., if threatened by the murderer)?
- Is the maxim of telling lies universalizable?
- Is it defeasible?
- Its it Ok to have a robot that tells lies:
  - What about Asimov Liar
  - What about HAL in



# Hypothetical imperatives

- Hypothetical imperative: they require us to do what fits our goals
  - I would like to have more money
  - If cheat on taxes I will have more money
  - I shall cheat on taxes to have more money
  
- I would like to get a good mark
  - If I study I will get a good mark
  - I shall study
  
- Is this OK?
- The imperative is dependent on what I want (getting good marks, having more money)
  - I shall cheat on taxes, to having more money!

# The categorical imperative

- A moral imperative that applies to all rational beings, irrespective of their personal wants and desires,
- “Act only on that maxim through which you can at the same time will that it should become a universal law”
  - - make false premises when it suits you to do so?
  - - refuse help to do those who are in need when it suits you to do so?



# The good will

- The morality of an action only depends only to the extent that this action is motivate by our good will, i.e., by the necessity to comply with the categorical imperative
  - E.g., if I do well my job only in order to get a promotion, or be better paid I am not acting morally
  - I am acting morally if I do well my job because I think that this is my categorical duty, since I believe that everybody should act upon the maxim that they ought to do well their job to ensure societal progress
- The good will is the only thing that is good in itself
  - Do you agree?

# Another version of the categorical imperative: the principle of humanity

- So act that you treat humanity in your own person and in the person of everyone else always at the same time as an end and never merely as means
  - How is it linked to universalizability: As you consider your self as an end, you should consider the others in the same way (universalizability)?
- What does it mean treating somebody as an end (not as a mere means)
  - It cannot mean that we never use people for our purposes (e.g., when we ask for favours or pay for jobs)
  - It must mean that we should never treat people ONLY as means, without considering their values and purposes

# When does AI treat people only as means

- Autonomous weapons?
- Deceiving advertisements?
- Discriminatory appointments?
  
- When does AI fail to recognise humans as valuable entities, that should achieve their aims according to their choices?
  
- Can we treat AI systems only as means?

# Dignity

- For Kant rational beings, capable of morality (humans) have a special status “an intrinsic worth, i.e., **dignity**,” which makes them valuable' “above all price
  - Because of dignity they deserve respect
  - They cannot be treated as mere ends
- What does it mean that AI systems should respect human dignity, respect humans



# The foundations of dignity

- Why do humans deserve dignity. Because they have
  - Reason: they act on reasons and are aware of this
  - Autonomy: they can choose what to do, and in particular to follow the categorical imperative rather than their subjective preference
- 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
- What if AI system also had reason and autonomy
- Would they become citizens of the kingdom of ends

# Morality as an aspect of rationality

- For Kant if we follow rationality, we have to be moral.
  - Can there be a rational criminal?
  - It is rational to pursue my wellbeing at the expense of others?
  - Is it rational for a company to develop a system that is profitable, but that will cause more harm than good (e.g.,

# Rationality and consistency

- 1. If you are rational, then you are consistent.
- 2. If you are consistent, then you obey the principle of universalizability.
- 3. If you obey the principle of universalizability, then you act morally.
- 4. Therefore, if you are rational, then you act morally.
- 5. Therefore, if you act immorally, then you are irrational.

What kind of consistency is this?

- If I deserve something no less than others, and I want it for me, I should recognise it also to others!
- Is this consistent with rationality? Is it required by it? Can I be rational, and pursue my goal to the detriment of other

# Issues

- Does the principle of universalizability always provide acceptable outcomes
- Is it sufficient that the maxim of my action is such that I would like it to be universalised for this maxim to be good?
- Can you think of some examples when this is not the case?
  - Lying ? Robbing? Celibacy? Genocide?



# Alan Gewirth (1912-2004): principle of generic consistency

1. I do (or intend to do) X voluntarily for a purpose E that I have chosen.
2. E is good
3. There are generic needs of agency.
4. My having the generic needs is good *for* my achieving E *whatever E might be*  $\equiv$  My having the generic needs is categorically instrumentally good for me.<sup>13</sup>
5. I categorically instrumentally ought to pursue my having the generic needs.
6. Other agents categorically ought not to interfere with my having the generic needs *against my will*, and ought to aid me to secure the generic needs when I cannot do so by my own unaided efforts *if I so wish*,
7. I am an agent  $\rightarrow$  I have the generic rights.
8. All agents have the generic rights.

Other attempts exist to develop a Kantian ethics.

# Approaches to universalisability

- Richard Hare (1919-2002)

- Moral judgments are universalizable: the judgment that an action is morally right/wrong commits me to accept that all relevantly similar actions are wrong
- Moral judgments are universalizable in the sense that they take into account the satisfaction of everybody's preferences (back to utilitarianism)

## Christine Korsgaard (1952)

- My humanity (capacity to reflectively act from reasons) is to me a source of value, and
- I must regard the humanity of others in the same way.

# 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

# David Ross (1877 1971): prima facie duties

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



# Defeasibility of duties

- Does it make sense to view duties as being defeasible?
- Can we apply defeasible reasoning to reason with duties?
- Should an AI system admit exceptions to duties, or should it always ask humans?

# Nietzsche(1844-1900) a critique of ethics

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

# Contractarianism

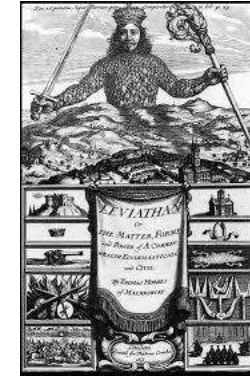
Giovanni Sartor

# 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)



# State of nature and social contract

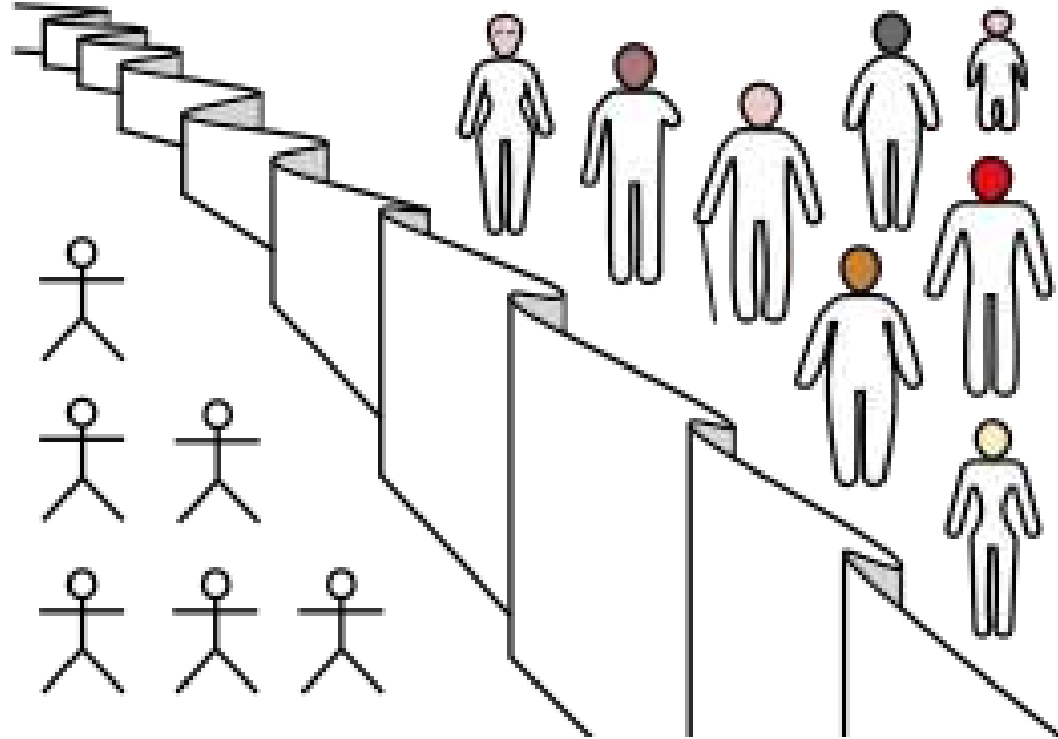


- How to get out of the state of nature?
- What agreements are OK?

		B	
		Cooperate	Defect
A	Cooperate	4, 4	-2, 6
	Defect	6, -2	0, 0

# John Rawls (1921-2002)

- A theory of justice
- How to ensure that the social contract is fair?
- People should choose under a **veil of ignorance**, without knowing their gender, social position, interests talents, wealth, race, etc.



# What principles would they go for?

- **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*). (JF, 42–43)

# AI in a just society (according to Rawls)

- Does the deployment of AI in today's society fit Rawls' requirements
- When may it conflict with the basic liberties?
- When with fair equality of opportunity?
- When with the difference principle?



# Juergen Habermas: Discourse Ethics

- A rule of action or choice is justified, and thus valid, only if all those affected by the rule or choice could accept it in a reasonable discourse.
- A norm is valid when the foreseeable consequences and side effects of its general observance for the interests and value orientations of each individual could be jointly accepted by all concerned without coercion
- The valid norms are those that would be the accepted outcome of an "ideal speech situation", in which all participants would be motivated solely by the desire to obtain a rational consensus and would evaluate each other's assertions solely on the basis of reason and evidence, being free of any physical and psychological coercion
- This approach assumes that people are able to engage in discourse and converge on the recognition of reasons for norms and choices

# Habermas and AI

- Would we all agree if we engaged in an impartial discussion on how to use AI?
- Can we think of an AI system that engages in an impartial moral debate? What would it argue for?

# Virtue ethics

Giovanni Sartor

# Virtue ethics

- Ethics should not focus on norms nor on consequences
  - An act is morally right just because it is one that a virtuous person, acting in character, would do in that situation.
- Ethics is a complex matter
  - Since there are many virtues, the right act is that that would result from the mix of the relevant virtues: honesty; loyalty; courage; impartiality, wisdom, fidelity, generosity, compassion, etc.
- Ethics cannot be learned through a set of rules, its application requires practical wisdom

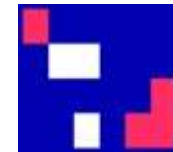
# Issues

- How do we know what is virtues and what is not?
- How can we extract precise indications from an account of virtues and from virtuous examples? How much can we rely in tradition?
- What if virtues are in conflict?
- What are the paradigms of virtues to which we may refer to?



# AI and virtue ethics

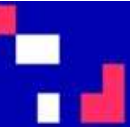
- Should we, as developer of AI systems, be virtuous? What character traits should we cultivate in us?
- Should AI applications (AI agents be virtuous)?
- How can virtues be learned?
- If from example, can the training of an AI system lead to a virtuous behaviour of it?



# Readings

- Shafer-Landau, R. (2018). The Fundamentals of Ethics. Oxford University Press.
- Singer, P. (2021). Ethics. In Encyclopedia Britannica:  
<https://www.britannica.com/topic/ethics-philosophy>





# Human Right and Information Technologies

Giovanni Sartor



Co-financed by the European Union  
Connecting Europe Facility

This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423



We are in a  
perilous  
navigation,  
since the ICT  
revolution

Turner: Dutch Fishing Boats in a Storm





Great  
opportunities





Great risks





At a crossroad  
between



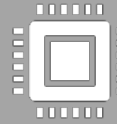
Multiple futures



# How to plan ahead?



Hard science: how things are ...



Technology: what is available/possible ...



Social science: what if ...



Normative knowledge: what values, what norms ...





# What normative knowledge

- General ethical theory, theories, computer ethics, machine ethics, AI ethics
- Regulations: data protection, consumer protection, competition law, civil liability, ...
- Human/fundamental rights and social values: the necessary link?

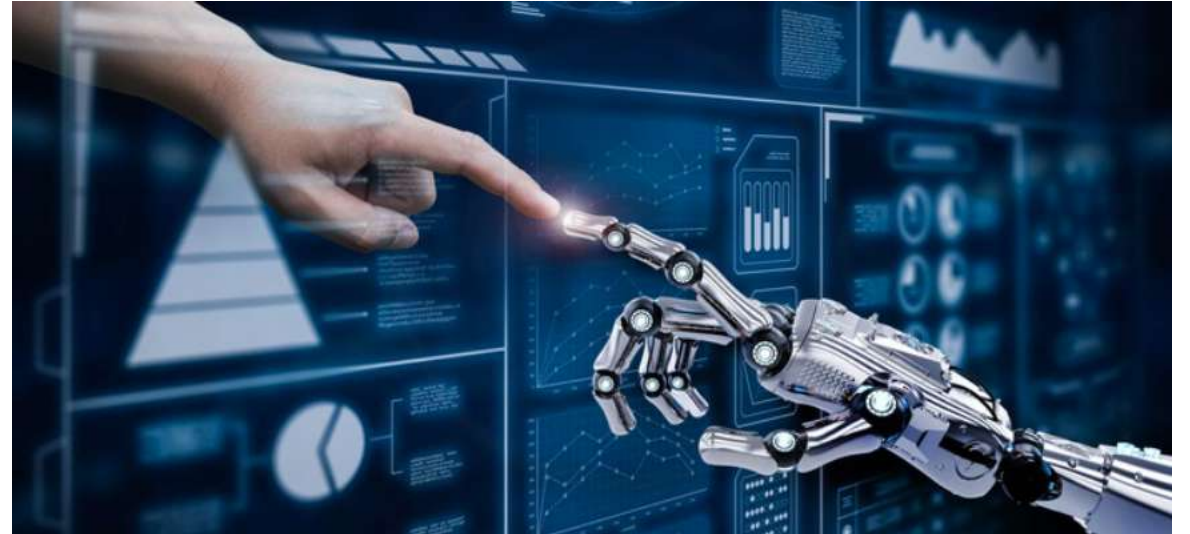


# AI4 People

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

# Trustworthy AI

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability



# A broader perspective: human values?



Human rights?





# A broad notion of human rights

- Primarily ethical demands (not to be “juridically incarcerated”)
- concerning freedoms (opportunities, including liberty and social rights) satisfying some “threshold conditions” of
  - special importance and
  - social influenceability.
- They may lead to
  - Imperfect duty (obligation to advocate, balance, take into account)
  - Perfect duties
- They may be the object of advocacy, of political debate, and (though not always) al legal enforcement

# ICT and human rights

- ITCs can
  - interfere with human rights,
  - Contribute to protect/implement human rights
  - provide for the existence of new human rights or add new content of existing right by
    - endowing a certain human opportunity with importance and
    - enabling society to realise it.
      - E.G.: right to access the internet, right to basic income, right to new medical technologies, etc.
- Not only an endangered legacy
- But also a blueprint for the future



Human/fundamental  
rights

As ethical  
rights



As political  
rights



As legal rights



# Human right in the big picture

- a future-oriented approach
- Human rights and an aspect of good ICT-pervaded society



# 1. Freedom and dignity

- All human beings are born free and equal in dignity and rights..



Pictures by Yacine Ait Kaci, from UDHR, UN 2015

## 7. Right to equality and nondiscrimination

- All are equal before the law and are entitled without any discrimination to equal protection of the law.
- All are entitled to equal protection against any discrimination .... and against any incitement to such discrimination.



## 12. Right to Privacy

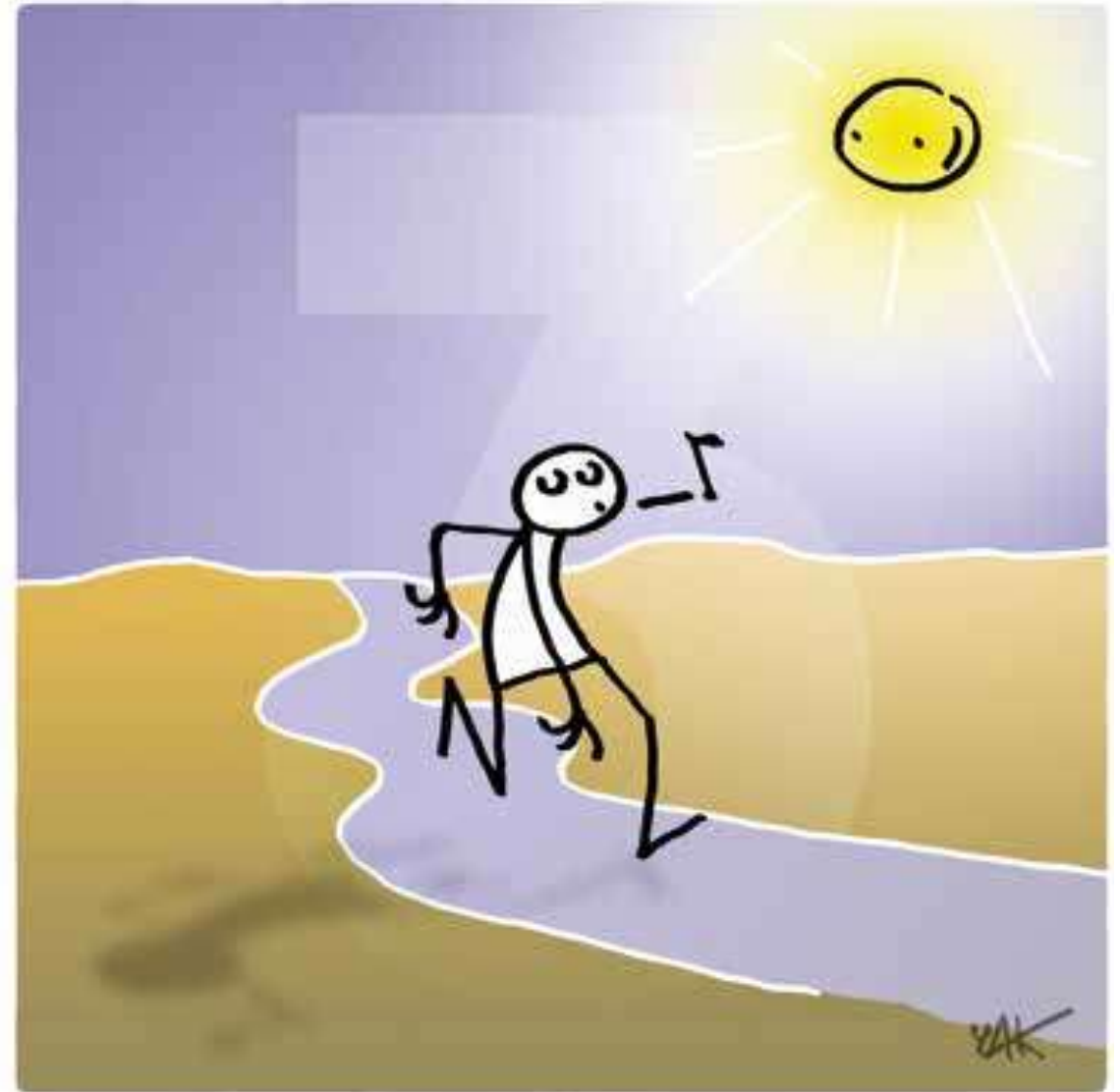
No one shall be subjected to arbitrary interference with his privacy, family, home or correspondence, nor to attacks upon his honour and reputation. Everyone has the right to the protection of the law against such interference or attacks





### 3. Right to life, liberty and security

- Everyone has the right to life, liberty and security of person.



Pictures by Yacine Ait Kaci, from UDHU, UN 2015

# 17. Right to property

- **(1) Everyone has the right to own property alone as well as in association with others.**



## 20. Freedom of assembly and association

- **(1) Everyone has the right to freedom of peaceful assembly and association.**
- **(2) No one may be compelled to belong to an association.**



## 8. Right to an effective remedy

- Everyone has the right to an effective remedy by the competent national tribunals for acts violating the fundamental rights granted him by the constitution or by law.



## 10. Right to a hearing

- Everyone is entitled in full equality to a fair and public hearing by an independent and impartial tribunal, in the determination of his rights and obligations and of any criminal charge against him.





# 11. Presumption of innocence

- **(1) Everyone charged with a penal offence has the right to be presumed innocent until proved guilty according to law in a public trial at which he has had all the guarantees necessary for his defence.**



# 19. Freedom of opinion, expression and information

Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers.



## 21. Right to take part in government

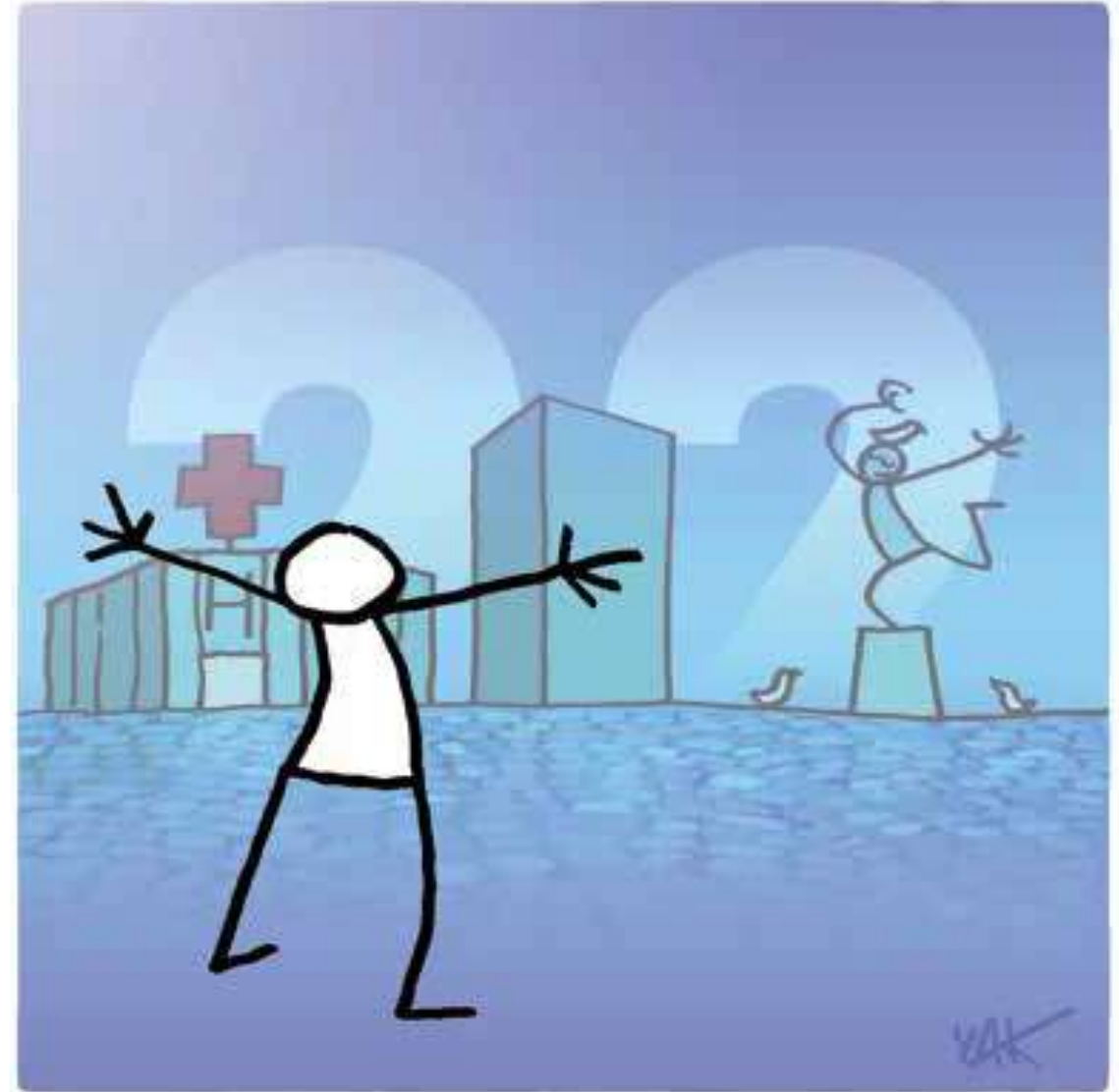
**(1) Everyone has the right to take part in the government of his country, directly or through freely chosen representatives.**

**(2) Everyone has the right to equal access to public service in his country.**



## 22. Right to social security

Everyone, as a member of society, has the right to social security and is entitled to realization [...] of the economic, social and cultural rights indispensable for his dignity and the free development of his personality.



## 23. Right to work

- **(1) Everyone has the right to work, to free choice of employment, to just and favourable conditions of work and to protection against unemployment**





## 25. Right to an adequate standard of living

- **(1) 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.**



## 26. Right to education

- **(1) Everyone has the right to education. Education shall be free, at least in the elementary and fundamental stages.**

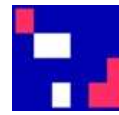


## 26. Right to culture

- **(1) 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.**



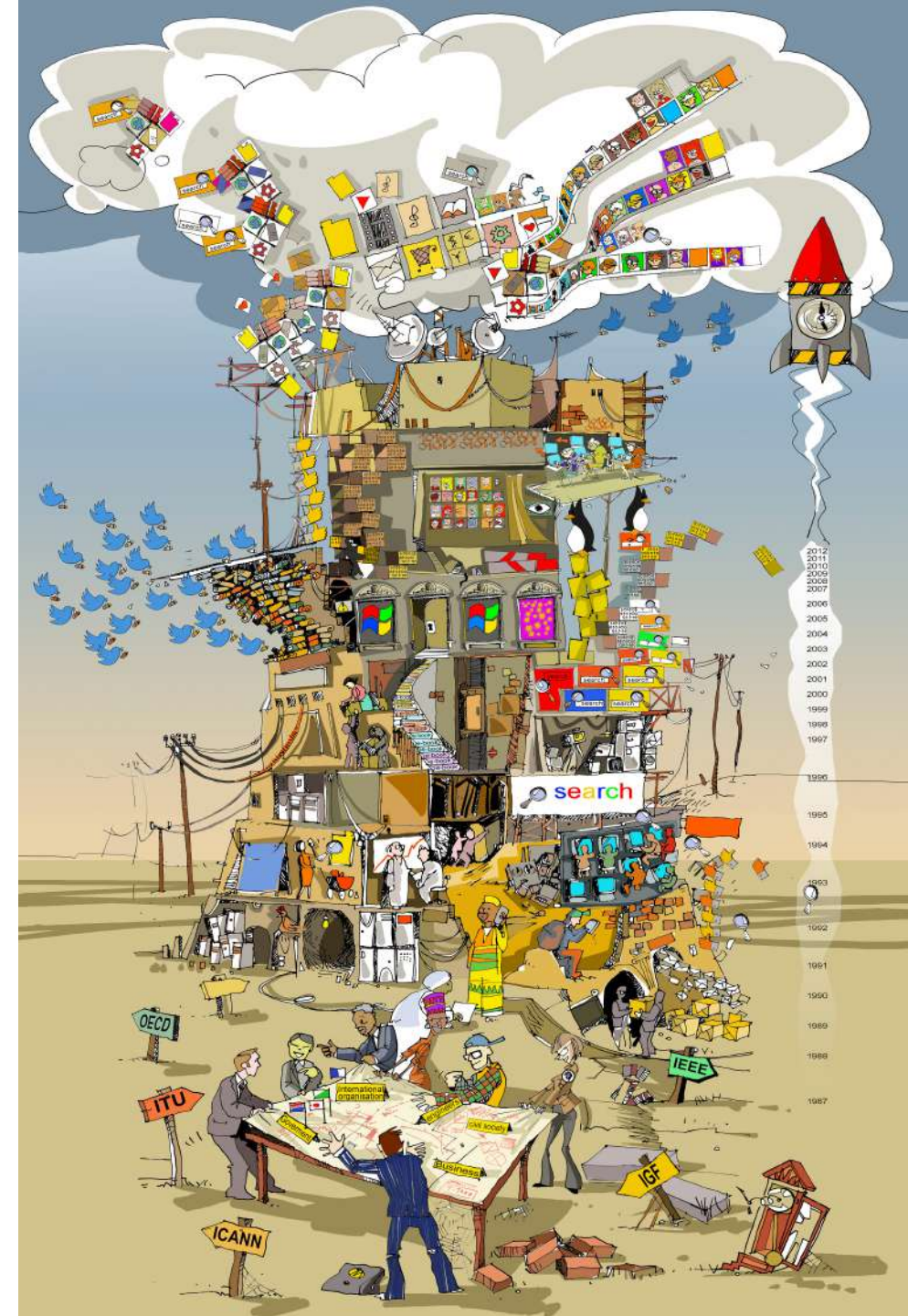


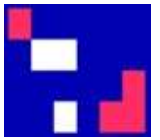


# Conclusion

- Human rights, as we have the ICT revolution are
  - A precious heritage to be protection, but also
  - blueprints for a human centred ict, and in particular human centred AI.

Thank for your attention  
[giovanni.sartor@eui.eu](mailto:giovanni.sartor@eui.eu)





# Logical English as a Programming Language for Law and Ethics

## A more human-friendly computer language for the future

Robert Kowalski with  
Jacinto Davila  
Galileo Sartor



Co-financed by the European Union  
Connecting Europe Facility



This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423





# A more human-friendly computer language for the future

## Logical English

- syntactic sugar for pure Prolog
- inspired in part by the language of well-written legal texts
- readable without any technical training in logic, computing or mathematics
- explainable
- incorporating deontic and other modalities
- not necessarily easy to write.

# Prolog in many natural languages

<https://legalmachinelab.unibo.it/logicalenglish/p/subset.pl>

<https://legalmachinelab.unibo.it/logicalenglish/p/subset-prolog.pl>

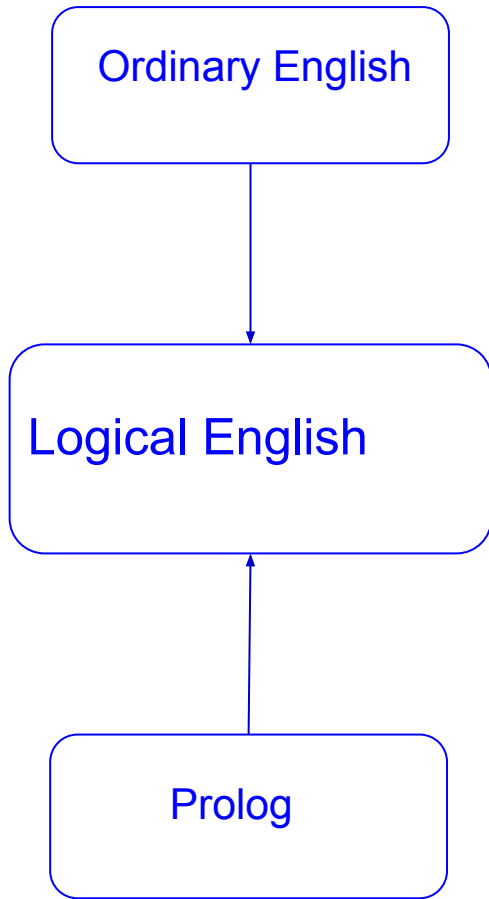
<https://legalmachinelab.unibo.it/logicalenglish/p/sousensemble.pl>

<https://legalmachinelab.unibo.it/logicalenglish/p/subconjunto.pl>

[https://legalmachinelab.unibo.it/logicalenglish/p/cittadinanza\\_ita.pl](https://legalmachinelab.unibo.it/logicalenglish/p/cittadinanza_ita.pl)

[https://legalmachinelab.unibo.it/logicalenglish/p/cittadinanza\\_ita-scasp.pl](https://legalmachinelab.unibo.it/logicalenglish/p/cittadinanza_ita-scasp.pl)

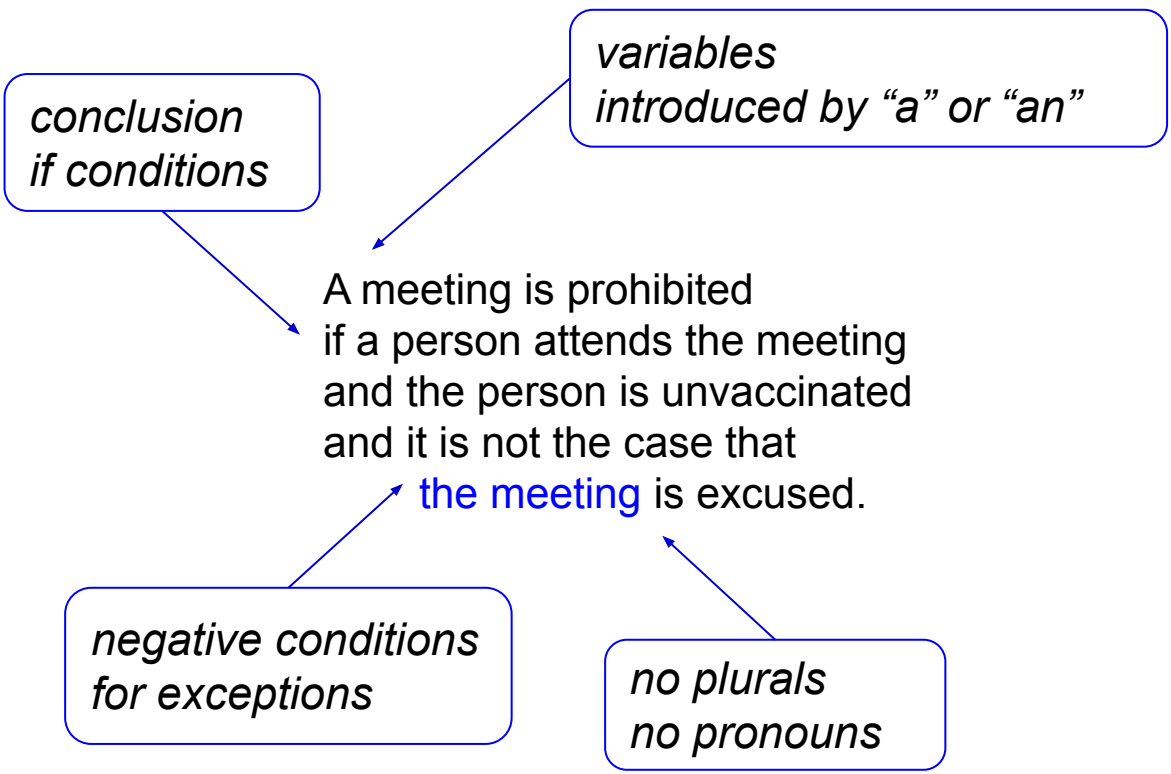
<https://legalmachinelab.unibo.it/logicalenglish/p/family-scasp.pl>



All meetings with unvaccinated people are prohibited unless **they** are excused.

A meeting is prohibited if a person attends the meeting and the person is unvaccinated and it is not the case that **the meeting** is excused.

```
prohibited(M)  
:- attends(P, M),  
   unvaccinated(P),  
   not(excused(M)).
```



Deontic modalities (obligation, prohibition, permission) can be represented by meta (or higher-order) predicates

A person has an obligation **that** the person pays an amount  
if the person attends a meeting  
and the meeting is prohibited  
and the fine for the person attending the meeting is the amount.

```
obligation(P, pays(P, A))  
:- attends(P, M),  
   prohibited(M),  
   fine(P, M, A).
```



Flood, M.D. and Goodenough, O.R., 2021. Contract as automaton: representing a simple financial agreement in computational form. *Artificial Intelligence and Law*, pp.1-26.

“Although the deontic approach reifies a number of key normative features, many expressions of formal law, such as statutes, regulations and the private rules of contract, typically do not use these normative formalisms in their natural language expressions.

Rather, such formal statements of law often substitute expressions of event and consequence for statements of obligation. That is, if certain rules are not respected, certain results—often unpleasant— will ensue.”

Deontic modalities (obligation, prohibition, permission) can be represented by specifying their consequences

A person has an obligation **that** the person pays an amount  
if the person attends a meeting  
and the meeting is prohibited  
and the fine for the person attending the meeting is the amount.

An arrest warrant is issued for a person  
if the person has an obligation that the person pays an amount  
and it is not the case that  
    the person pays the amount.

To be fully operational, the payment and arrest warrant events would need explicit temporal constraints.

# Logical English on SWISH (online version of SWI Prolog)

```
19 A meeting is prohibited
20   if a person attends the meeting
21   and the person is unvaccinated
22   and it is not the case that
23     the meeting is excused.
24
25 A person has an obligation that the person pays an amount
26   if the person attends a meeting
27   and the meeting is prohibited
28   and the fine for the person attending the meeting is the amount.
29
30 An arrest warrant is issued for a person
31   if the person has an obligation that the person pays an amount
32   and it is not the case that
33     the person pays the amount.
34
35 scenario one is:
36
37   Boris attends christmas party.
38   Novak attends christmas party.
39   Novak is unvaccinated.
40   the fine for a person attending a meeting is £100
41   if the meeting is prohibited.
42 %   Novak pays £100.
43   Boris pays £1000.
```

```
answer("query one with scenario one").
```

```
Query one with one: *a person* pays *an amount*
```

```
Answer: Boris pays £ 1000
```

```
true
```

```
answer("query two with scenario one").
```

```
Query two with one: An arrest warrant is issued for *a person*
```

```
Answer: An arrest warrant is issued for Boris
```

```
true
```

```
Answer: An arrest warrant is issued for Novak
```

```
true
```

```
?- answer("query two with scenario one").
```

Examples▲

History▲

Solutions▲

# The British Nationality Act

18 a person acquires British citizenship on a date  
19 if the person is born in the UK on the date  
20 and the date is after commencement  
21 and an other person is the mother of the person  
22 or the other person is the father of the person  
23 and the other person is a British citizen on the date  
24 or the other person is settled in the UK on the date,  
25

```
acquires_British_citizenship_on(A, B) :-  
    is_born_in_on(A, the_UK, B),  
    is_after_commencement(B),  
    (  
        is_the_mother_of(C, A)  
    ;   is_the_father_of(C, A)  
    ),  
    (  
        is_a_British_citizen_on(C, B)  
    ;   is_settled_in_the_UK_on(C, B)  
    ).
```

?-

show prolog.

Examples▲

History▲

Solutions▲

table results

Run!



```
67
68 scenario harry is:
69
70 John is born in the UK on 2021-10-09.
71 2021-10-09 is after commencement.
72 Harry is the father of John.
73 Harry is settled in the UK on 2021-10-09.
74
75 query one is:
76
77 which person acquires British citizenship on which date.
78
```

Answer: John acquires British citizenship on 2021-10-9T0:0:0.0

true

Next 10 100 1,000 Stop

?- answer one with harry.

Examples▲ History▲ Solutions▲

table results

Run!

```
67
68 scenario harry_says is:
69
70 John is born in the UK on 2021-10-09.
71 2021-10-09 is after commencement.
72 Harry says that Harry is the father of John.
73 Harry is settled in the UK on 2021-10-09.
74
75 query one is:
76
77 which person acquires British citizenship on which date.
78
```

The screenshot shows a Prolog IDE interface. At the top, a text box contains the word `false` in red. Below it, a query prompt `?-` is followed by a text box containing the query `answer one with harry_says.`. At the bottom of the interface, there are three buttons labeled `Examples▲`, `History▲`, and `Solutions▲`. To the right of these buttons is a checkbox labeled `table results` and a blue button labeled `Run!`.

```
54
55 scenario trust is:
56
57 John is born in the UK on 2021-10-09.
58 2021-10-09 is after commencement.
59 Harry says that Harry is the father of John.
60 Harry is settled in the UK on 2021-10-09.
61
62 a person X is the father of a person Y
63   if X says that X is the father of Y.
64
65 query one is:
66
67 which person acquires British citizenship on which date.
```

The screenshot shows a logic programming interface. At the top, a text box contains the answer: "Answer: John acquires British citizenship on 2021-10-9T0:0:0.0". Below this, the word "true" is displayed. A control bar contains buttons for "Next", "10", "100", "1,000", and "Stop". The main query area shows a query: "?- answer one with trust.", where "with" is highlighted with a dashed box. At the bottom, there are buttons for "Examples", "History", and "Solutions", a checkbox for "table results", and a blue "Run!" button.

15-04 | March 26, 2015  
*Revised March 27, 2017*

*Artificial Intelligence and Law, 2021*

## Contract as Automaton: The Computational Representation of Financial Agreements

**Mark D. Flood**

Office of Financial Research  
[mark.flood@ofr.treasury.gov](mailto:mark.flood@ofr.treasury.gov)

**Oliver R. Goodenough**

Office of Financial Research and Vermont Law School  
[oliver.goodenough@ofr.treasury.gov](mailto:oliver.goodenough@ofr.treasury.gov)  
[ogoodenough@vermontlaw.edu](mailto:ogoodenough@vermontlaw.edu)

**Table 2: A Streamlined Loan Agreement**

**Agreement**

This loan agreement dated June 1, 2014, by and between Lender Bank Co. ("Lender") and Borrower Corp. (Borrower), will set out the terms under which Lender will extend credit in the principal amount of \$1,000 to Borrower with an un-compounded interest rate of 5% per annum, included in the specified payment structure.

**1. The Loan:**

At the request of Borrower, to be given on June 1, 2014, Lender will advance \$1000 to Borrower no later than June 2, 2014. If Borrower does not make such a request, this agreement will terminate.

**2. Repayment:**

Subject to the other terms of this agreement, Borrower will repay the loan in the following payments:

- (a) Payment 1, due June 1, 2015, in the amount of \$550, representing a payment of \$500 as half of the principal and interest in the amount of \$50.
- (b) Payment 2, due June 1, 2016, in the amount of \$525, representing a payment of \$500 as the remaining half of the principal and interest in the amount of \$25.

**3. Representations and Warranties:**

The Borrower represents and warrants, at the execution of this agreement, at the request for the advance of funds and at all times any repayment amount shall be outstanding, the Borrower's assets shall exceed its liabilities as determined under an application of the FASB rules of accounting.

**4. Covenants:**

The Borrower covenants that at the execution of this agreement, at the request for the advance of funds and at all times any repayment amount shall be outstanding it will make timely payment of all state and federal taxes as and when due.

**5. Events of Default:**

The Borrower will be in default under this agreement upon the occurrence of any of the following events or conditions, provided they shall remain uncured within a period of two days after notice is given to Borrower by Lender of their occurrence (such an uncured event an "Event of Default"):

- (a) Borrower shall fail to make timely payment of any amount due to Lender hereunder;
- (b) Any of the representation or warranties of Borrower under this agreement shall prove untrue;
- (c) Borrower shall fail to perform any of its covenants under this agreement;
- (d) Borrower shall file for bankruptcy or insolvency under any applicable federal or state law.

A default will be cured by the Borrower (i) remedying the potential event of default and (ii) giving effective notice of such remedy to the Lender. In the event of multiple events of default, the first

to occur shall take precedence for the purposes of specifying outcomes under this agreement.

**6. Acceleration on Default**

Upon the occurrence of an Event of Default all outstanding payments under this agreement will become immediately due and payable, including both principal and interest amounts, without further notice, presentment, or demand to the Borrower.

**7. Choice of Law:**

This agreement will be subject to the laws of the State of New York applicable to contracts entered into and performed wholly within that state.

**8. Amendments and Waivers:**

Any purported amendment to, or waiver of rights under, this agreement will only be effective if set forth in writing and executed by both parties.

**9. Courts and Litigation:**

Any legal action brought to enforce, interpret or otherwise deal with this agreement must be brought in the state courts of the State of New York located in New York County, and each of the parties agrees to the jurisdiction of such courts over both the parties themselves and over the subject matter of such a proceeding, and waives any claim that such a court may be an inconvenient forum.

**10. Time of the Essence; No Pre-Payment**

Timely performance is required for any action to be taken under this agreement, and, except as may otherwise be specifically provided herein, failure to take such action on the day specified will constitute a binding failure to take such action. Payments shall only be made on or after the dates specified in Section 2 or on or after such other date as may be required under Section 6; pre-payments made on earlier dates shall not be accepted.

**11. Notices**

Notices provided for in this agreement will be given by an email to the email addresses set out below and will be effective upon receipt.

[Lender email here]

[Borrower email here]

Accepted and agreed:

LENDER BANK CO.

BORROWER CORP.

By: \_\_\_\_\_

By: \_\_\_\_\_

Title: \_\_\_\_\_

Title: \_\_\_\_\_

[NOTE: Statute of Limitations on debt obligations in NY is 6 years]

Draft of July 23, 2014







## When does an Event of Default occur?

### 5. Events of Default:

on day D0

The Borrower will be in default under this agreement upon the occurrence of any of the following events or conditions, provided they shall remain uncured within a period of two days after notice is given to Borrower by Lender of their occurrence (such an uncured event an “Event of Default”):

on day D0

- (a) Borrower shall fail to make timely payment of any amount due to Lender hereunder;
- (b) Any of the representation or warranties of Borrower under this agreement shall prove untrue;
- (c) Borrower shall fail to perform any of its covenants under this agreement;
- (d) Borrower shall file for bankruptcy or insolvency under any applicable federal or state law.

## When does an Event of Default occur?

### 5. Events of Default:

if

on day D0

The Borrower will be in default under this agreement upon the occurrence of any of the following events or conditions, provided they shall remain uncured within a period of two days after notice is given to Borrower by Lender of their occurrence (such an uncured event an “Event of Default”):

on day D1

on day D0

- (a) Borrower shall fail to make timely payment of any amount due to Lender hereunder;
- (b) Any of the representation or warranties of Borrower under this agreement shall prove untrue;
- (c) Borrower shall fail to perform any of its covenants under this agreement;
- (d) Borrower shall file for bankruptcy or insolvency under any applicable federal or state law.

## When does an Event of Default occur?

### 5. Events of Default:

if

on day D0

The Borrower will be in default under this agreement upon the occurrence of any of the following events or conditions, provided they shall remain uncured within a period of two days after notice is given to Borrower by Lender of their occurrence (such an uncured event an “Event of Default”):

on day D1

on day D0

- (a) Borrower shall fail to make timely payment of any amount due to Lender hereunder;
- (b) Any of the representation or warranties of Borrower under this agreement shall prove untrue;
- (c) Borrower shall fail to perform any of its covenants under this agreement;
- (d) Borrower shall file for bankruptcy or insolvency under any applicable federal or state law.

A default will be cured by the Borrower

- (i) remedying the potential event of default and
- (ii) giving effective notice of such remedy to the Lender.



## When does an Event of Default occur?

### 5. Events of Default:

if

on day D0

The Borrower will be in default under this agreement upon the occurrence of any of the following events or conditions, provided they shall remain uncured within a period of two days after notice is given to Borrower by Lender of their occurrence

(such an uncured event an “Event of Default”):

on D2 = D1 + 2

on day D1

on day D0

- (a) Borrower shall fail to make timely payment of any amount due to Lender hereunder;
- (b) Any of the representation or warranties of Borrower under this agreement shall prove untrue;
- (c) Borrower shall fail to perform any of its covenants under this agreement;
- (d) Borrower shall file for bankruptcy or insolvency under any applicable federal or state law.

A default will be cured by the Borrower

- (i) remedying the potential event of default and
- (ii) giving effective notice of such remedy to the Lender.

## When does an Event of Default occur?

### 5. Events of Default:

if

on day D0

The Borrower will be in default under this agreement upon the occurrence of any of the following events or conditions, provided they shall remain uncured within a period of two days after notice is given to Borrower by Lender of their occurrence

(such an uncured event an “Event of Default”):

on D2 = D1 + 2

on D0 or D2?

on day D1

on day D0

- (a) Borrower shall fail to make timely payment of any amount due to Lender hereunder;
- (b) Any of the representation or warranties of Borrower under this agreement shall prove untrue;
- (c) Borrower shall fail to perform any of its covenants under this agreement;
- (d) Borrower shall file for bankruptcy or insolvency under any applicable federal or state law.

6. Acceleration on Default. Upon the occurrence of an Event of Default all outstanding payments under this agreement will become immediately due and payable, including both principal and interest amounts, without further notice, presentment, or demand to the Borrower.

28 the borrower defaults on a date  $D_2$   
29 if the borrower has an obligation  
30 and the borrower fails on a date  $D_0$  to fulfil the obligation  
31 and the Lender notifies the borrower on a date  $D_1$   
32 that the borrower fails on  $D_0$  to fulfil the obligation  
33 and  $D_2$  is 2 days after  $D_1$   
34 and it is not the case that  
35 the borrower cures the failure of the obligation on or before  $D_2$ .

36  
37 the borrower cures the failure of an obligation on or before a date  $D_3$   
38 if the obligation is  
39 that the borrower pays an amount to the Lender on a date  $D_0$   
40 and the borrower pays the amount to the Lender on a date  $D_1$   
41 and the borrower notifies the Lender on a date  $D_2$   
42 that the borrower pays the amount to the Lender on  $D_1$   
43 and  $D_1$  is on or before  $D_3$   
44 and  $D_2$  is on or before  $D_3$ .

28 the borrower defaults on a date  $D_2$   
29 if the borrower has an obligation  
30 and the borrower fails on a date  $D_0$  to fulfil the obligation  
31 and the lender notifies the borrower on a date  $D_1$   
32 that the borrower fails on  $D_0$  to fulfil the obligation  
33 and  $D_2$  is 2 days after  $D_1$   
34 and it is not the case that  
35 the borrower cures the failure of the obligation on or before  $D_2$ .

36  
37 the borrower cures the failure of an obligation on or before a date  $D$   
38 if the obligation is  
39 that the borrower pays an amount to the lender on an other date  
40 and the borrower pays the amount to the lender on a new payment date  
41 and the borrower notifies the lender on a notification date  
42 that the borrower pays the amount to the lender on the new payment date  
43 and the new payment date is on or before  $D$   
44 and the notification date is on or before  $D$ .

52 scenario payment is:

53 the Lender notifies the borrower on 2016-06-04

54 that the borrower fails on 2016-06-01 to fulfil obligation2.

55 the borrower pays 525 to the Lender on 2016-06-05.

56 the borrower notifies the Lender on 2016-06-06

57 that the borrower pays 525 to the Lender on 2016-06-06.

58

59 query defaults is:

60 which person defaults on which day.

Query defaults with payment: \*a borrower\* defaults on \*a date\*

Answer: the borrower defaults on 2016-6-6T0:0:0.0

true

?- answer defaults with payment.



# Prospects for the Future

- All computer languages should be readable without training.
- But learning to write will be harder than learning to read.
- Learning to write well will be much harder.
- We need a corpus of well-written examples.
- Legal applications are a good place to start.

SWISH implementation of LE at

<https://logicalenglish.logicalcontracts.com/>

Some other examples

```

1 :- module('subset+http://tests.com', []).
2
3 en("the target language is: prolog.
4
5 the templates are:
6   *a set* is a subset of *a set*,
7   *a thing* is a set,
8   *a thing* belongs to *a set*.
9
10 the knowledge base subset includes:
11
12 a set A is a subset of a set B
13   if set A is a set
14   and set B is a set
15   and for all cases in which
16   a thing belongs to set A
17   it is the case that
18   the thing belongs to set B.
19
20 scenario one is:
21   family one is a set.
22   family two is a set.
23   Bob belongs to family one.
24   Alice belongs to family one.
25
26   Alice belongs to family two.
27
28 query one is:
29   which first family is a subset of which second family.
30
31 scenario two is:
32   [Alice, Bob] is a set.
33   [Alice] is a set.
34
35   a thing belongs to a set

```

answer one with one.

Query one with one: \*a set\* is a subset of \*a set\*

Answer: family one is a subset of family one

true

Answer: family two is a subset of family one

true

Answer: family two is a subset of family two

true

---

answer two with two.

Query two with two: \*a set\* is a subset of \*a set\*

Answer: [Alice, Bob] is a subset of [Alice, Bob]

true

Next 10 100 1,000 Stop

---

show prolog.

```

is_a_subset_of(A, B) :-
  is_a_set(A),
  is_a_set(B),
  forall(belongs_to(C, A), belongs_to(C, B)).
query(null, true).
query(one, is_a_subset_of(_, _)).
query(two, is_a_subset_of(_, _)).
example(null, []).
example(one, [scenario([is_a_set(family_one):-true), (is_a_set(family_two):-true), (belongs_to('Bob', family_one):-true), (belongs_to('Alice', family_one):-true), (belongs_to('Alice', family_two):-true)]]).
example(two, [scenario([is_a_set(['Alice', 'Bob']):-true), (is_a_set(['Alice']):-true), (belongs_to('Alice', family_one):-true), (belongs_to('Alice', family_two):-true)]]).
true

```

show prolog.

Examples History Solutions

table results Run!

```
11
12 un ensemble A est un sous-ensemble d'un ensemble B
13   si L'ensemble A est un ensemble
14   et L'ensemble B est un ensemble
15   et pour tous Les cas où
16     une chose appartient à L'ensemble A
17     c'est le cas que
18     La chose appartient à L'ensemble B.
19
20 Le scénario un est:
21   La famille un est un ensemble.
22   La famille deux est un ensemble.
23   Bob appartient à La famille un.
24   Alice appartient à La famille un.
25
26   Alice appartient à La famille deux.
27
28 La question un est:
29   quelle premier famille est un sous-ensemble d' quelle seconde famille.
30
31 Le scénario deux est:
32   [Alice, Bob] est un ensemble.
33   [Alice] est un ensemble.
34
```

répondre deux avec un.

La question deux avec un: \*an ensemble\* est un sous-ensemble d \*an ensemble\*

La réponse: La famille un est un sous-ensemble d La famille un

true 1

La réponse: La famille deux est un sous-ensemble d La famille un

true 2

La réponse: La famille deux est un sous-ensemble d La famille deux

true 3

répondre deux avec deux.

La question deux avec deux: \*an ensemble\* est un sous-ensemble d \*an ensemble\*

La réponse: [Alice,Bob] est un sous-ensemble d [Alice,Bob]

true 1

La réponse: [Alice] est un sous-ensemble d [Alice,Bob]

true 2

La réponse: [Alice] est un sous-ensemble d [Alice]

true 3

?- répondre deux avec deux.

Examples History Solutions  table results Run!

12 un conjunto A es un subconjunto de un conjunto B  
13 si el conjunto A es un conjunto  
14 y el conjunto B es un conjunto  
15 y for all cases in which  
16 una cosa pertenece a el conjunto A  
17 it is the case that  
18 La cosa pertenece a el conjunto B.  
19  
20 escenario uno es:  
21 familia uno es un conjunto.  
22 familia dos es un conjunto.  
23 Roberto pertenece a la familia uno.  
24 Alicia pertenece a la familia uno.  
25  
26 Alicia pertenece a la familia dos.  
27  
28 La pregunta uno es:  
29 which first familia es un subconjunto de which second familia.  
30  
31 escenario dos es:  
32 [Alicia, Roberto] es un conjunto.  
33 [Alicia] es un conjunto.  
34  
35 una cosa pertenece a un conjunto  
36 if la cosa is in el conjunto.  
37  
38 La pregunta dos es:  
39 which conjunto es un subconjunto de which other conjunto.  
40

responde dos con uno.

La pregunta dos con uno: \*a conjunto\* es un subconjunto de \*a conjunto\*

La respuesta: familia uno es un subconjunto de familia uno

true

La respuesta: familia uno es un subconjunto de familia dos

true

La respuesta: familia dos es un subconjunto de familia uno

true

La respuesta: familia dos es un subconjunto de familia dos

true

responde dos con dos.

La pregunta dos con dos: \*a conjunto\* es un subconjunto de \*a conjunto\*

La respuesta: [Alicia,Roberto] es un subconjunto de [Alicia,Roberto]

true

La respuesta: [Alicia] es un subconjunto de [Alicia,Roberto]

true

La respuesta: [Alicia] es un subconjunto de [Alicia]

true

?- responde dos con dos.

Examples▲

History▲

Solutions▲



16 La base di conoscenza cittadinanzaaita include:  
17  
18 una persona A ha la cittadinanza italiana  
19 se una persona B è genitore di La persona A  
20 e La persona B ha la cittadinanza italiana.  
21  
22 una persona A è genitore di una persona B  
23 se La persona A è padre di La persona B.  
24  
25 una persona A è genitore di una persona B  
26 se La persona A è madre di La persona B.  
27  
28 scenario giuseppe è:  
29 felice è padre di giuseppe.  
30 tatiana è madre di giuseppe.  
31 felice ha la cittadinanza italiana.  
32 tatiana ha la cittadinanza italiana.  
33  
34 domanda uno è:  
35 quale persona ha la cittadinanza italiana.  
36

 answer(uno, with(giuseppe), le(E), R).

E =

It is the case that: **giuseppe ha la cittadinanza italiana** as proved by [KB Text](#)  
because

- It is the case that: **felice è genitore di giuseppe** as proved by [KB Text](#)  
because
  - It is the case that: **felice è padre di giuseppe** as proved by *hypothesis in scenario*
- It is the case that: **felice ha la cittadinanza italiana** as proved by *hypothesis in scenario*,

R = true

E =

It is the case that: **giuseppe ha la cittadinanza italiana** as proved by [KB Text](#)  
because

- It is the case that: **tatiana è genitore di giuseppe** as proved by [KB Text](#)  
because
  - It is the case that: **tatiana è madre di giuseppe** as proved by *hypothesis in scenario*
- It is the case that: **tatiana ha la cittadinanza italiana** as proved by *hypothesis in scenario*,

R = true

Next 10 100 1 000 Stop  
Examples History Solutions

ta

# The event calculus for reasoning about time

14 a fluent holds at a time  $T_2$   
15 if an event happens at a time  $T_1$   
16 and the event initiates the fluent at  $T_1$   
17 and  $T_1$  is before  $T_2$   
18 and it is not the case that  
19 an other event happens at a time  $T$   
20 and the other event terminates the fluent at  $T$   
21 and  $T_1$  is on or before  $T$   
22 and  $T$  is before  $T_2$ .

23  
24 switch up initiates light at each time.  
25 switch down initiates dark at each time.  
26 switch up terminates dark at each time.  
27 switch down terminates light at each time.

28  
29 switch up happens at 1.  
30 switch down happens at 4.

Answer: light holds at 2

true

Answer: light holds at 3

true

Answer: light holds at 4

true

Answer: dark holds at 5

true

Answer: dark holds at 6

true

Next

10

100

1,000

Stop

?- answer two with switching.

Examples▲

History▲

Solutions▲

```
10
11 the knowledge base simple RPS includes:
12     scissors beats paper.
13     paper beats rock.
14     rock beats scissors.
15
16 a first player gets a prize
17     if the first player inputs a first choice and an amount X
18     and a second player inputs a second choice and an amount Y
19     and the first player is different from the second player
20     and the first choice beats the second choice
21     and the prize is X+Y.
22
23 the game is a draw
24     if a first player inputs a first choice and an amount X
25     and a second player inputs the first choice and an amount Y
26     and the first player is different from the second player.
27
28 a player gets an amount
29     if the game is a draw
30     and the player inputs a choice and the amount.
31
32 scenario mb is:
33     miguel inputs scissors and 100.
34     bob inputs paper and 1000.
35
36 query gets is:
37     which person gets which amount.
```

answer gets with mb.

Query gets with mb: \*a person\* gets \*an amount\*

Answer: miguel gets 1100

true


false

?- answer gets with mb.

```

12 scissors beats paper.
13 paper beats rock.
14 rock beats scissors.
15
16 a first player gets a prize
17 if the first player inputs a first choice and an amount X
18 and a second player inputs a second choice and an amount Y
19 and the first player is different from the second player
20 and the first choice beats the second choice
21 and the prize is X+Y.
22
23 the game is a draw
24 if a first player inputs a first choice and an amount X
25 and a second player inputs the first choice and an amount Y
26 and the first player is different from the second player.
27
28 a player gets an amount
29 if the game is a draw
30 and the player inputs a choice and the amount.
31
32 scenario mbj is:
33 miguel inputs paper and 100.
34 bob inputs paper and 1000.
35 % jacinto inputs paper and 1000.
36
37 query gets is:
38 which person gets which amount.
39
40 query draw is:
41 the game is a draw.

```

 `answer(gets, with(mbj), le(Explanations), R).`

**Explanations =**

- It is the case that: **miguel gets 100** as proved by [KB Text](#) because
  - It is the case that: **the game is a draw** as proved by [KB Text](#) because
    - It is the case that: **miguel inputs paper and 100** as proved by *hypothesis in scenario*
    - It is the case that: **bob inputs paper and 1000** as proved by *hypothesis in scenario*

**R = true**

**Explanations =**

- It is the case that: **bob gets 1000** as proved by [KB Text](#) because
  - It is the case that: **the game is a draw** as proved by [KB Text](#) because
    - It is the case that: **miguel inputs paper and 100** as proved by *hypothesis in scenario*
    - It is the case that: **bob inputs paper and 1000** as proved by *hypothesis in scenario*

**R = true**

Next 10 100 1,000 Stop

---

?- `answer(gets, with(mbj), Le(Explanations), R).`

Examples History Solutions  table res



# Explainable and Ethical AI: A Perspective on Argumentation and Logic Programming

Ethics in Artificial Intelligence  
Technologies

Roberta Calegari

[roberta.calegari@unibo.it](mailto:roberta.calegari@unibo.it)

ALMA MATER STUDIORUM – Università di Bologna

Academic Year 2020/2021



Co-financed by the European Union  
Connecting Europe Facility

This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423





# Next in Line...

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
- 5 An LP approach to Ethics
- 6 Architecture
- 7 Discussion



# Context I

## Context of AI applications

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



## Context II

### Investigations on programming machine ethics

- one stressing above all individual cognition, deliberation, and behavior
  - computation is vehicle for the study of morality, namely in its modeling of the dynamics of knowledge and cognition of agents
  - addressing moral facets such as permissibility and the dual process of moral judgments by framing together various logic programming (LP) knowledge representation and reasoning features that are essential to moral agency
    - abduction with integrity constraints
    - preferences over abductive scenarios
    - probabilistic reasoning
    - counterfactuals, and updating
    - argumentation
- the other stressing collective morals, and how they emerged

## Context III

### LP and morality

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

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



# Next in Line...

- 1 Context
- 2 Agents**
- 3 Motivation
- 4 Preliminaries
- 5 An LP approach to Ethics
- 6 Architecture
- 7 Discussion





# Agents as Autonomous Entities (*recap*)

## Definition (Agent)

Agents are *autonomous computational entities* [Omicini et al., 2008]

**genus** agents are computational entities

**differentia** agents are autonomous, in that they encapsulate control along with a criterion to govern it

## Agents are *autonomous*

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

# Next in Line...

- 1 Context
- 2 Agents
- 3 Motivation**
- 4 Preliminaries
- 5 An LP approach to Ethics
- 6 Architecture
- 7 Discussion



# Why Logic? I

Logic-based approaches already play a well-understood role in the engineering of intelligent (multi-agent) systems; declarative, logic-based approaches have the potential to represent an alternative way of delivering symbolic intelligence, complementary to the one pursued by sub-symbolic approaches [Calegari et al., 2020].

- Logic-based technologies address opaqueness issues, and, once suitably integrated with argumentation capabilities, can provide for features like interpretability, observability, accountability, and explainability.
- well-founded definition of explanation (abducible, *conversation...*)

# Why Logic? II

## LP reasoning features

- *Abduction* scenario generation and of hypothetical reasoning, including the consideration of counterfactual scenarios about the past
- *Preferences* enacted for preferring scenarios obtained by abduction
- *Probabilistic LP* allows abduction to take scenario uncertainty into account
- *LP counterfactuals* permit hypothesizing into the past, even taking into account present knowledge
- *Argumentation* converse, debate and explain

And technically

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

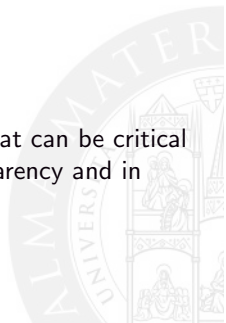
## Why Logic? III

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

The main answer lies in the three main features of LP

- (i) being a declarative paradigm
- (ii) working as a tool for knowledge representation, and
- (iii) allowing for different forms of reasoning and inference

These features lead to some properties for intelligent systems that can be critical in the design of ubiquitous intelligence (both in terms of transparency and in terms of ethics).





# Why Logic? IV

## Provability

- correctness, completeness, well-founded extension
- ensuring some fundamental computational properties – such as correctness and completeness.
- extensions can be formalised, well-founded as well, based on recognised theorems

Provability is a key feature in the case of trusted and safe systems.



# Why Logic? V

## Explainability

- formal methods for argumentation-, justification-, and counterfactual often based on LP [Saptawijaya and Pereira, 2019]
- system capable to engage in dialogues with other actors to communicate its reasoning, explain its choices, or to coordinate in the pursuit of a common goal
- other logical forms of explanation can be envisaged via non-monotonic reasoning and argumentation, through a direct extension of the semantics of LP



# Why Logic? VI

## Expressivity and situatedness

- different nuances → extensions [Dyckhoff et al., 1996]
- explicit assumptions and exceptions [Borning et al., 1989]
- capture the specificities of the context [Calegari et al., 2018b]



# Why Logic? VII

## Hybridization

- integration of diversity [Calegari et al., 2018a]
- represent the heterogeneity of the contexts of intelligent systems – also in relation to the application domains – and to customise as needed the symbolic intelligence that is provided while remaining within a well-founded formal framework



# Why Logic for Agents?

- it is a declarative, logic programming language, yet not an agent programming language
  - with a built-in control mechanism, not a theory of agency
- logic inference for reasoning
- reasoning for deliberation
- explicit belief and goal representation for agent-oriented operations
- could be used to build cognitional artefacts





# Next in Line...

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries**
- 5 An LP approach to Ethics
- 6 Architecture
- 7 Discussion



# Essentials of LP I

## Three fundamental features [Apt, 2005]

**terms** *Computing takes place over the domain of all terms defined over a “universal” alphabet.*

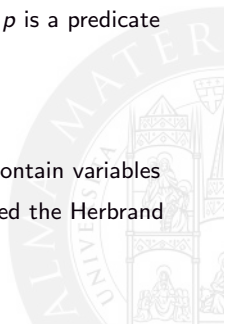
**mgu** *Values are assigned to variables by means of automatically-generated substitutions, called most general unifiers. These values may contain variables, called logical variables.*

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



# Essentials of LP II

- Let  $A$  be an alphabet of a language  $L$
- countable disjoint set of constants, function symbols, and predicate symbols.
- an alphabet is assumed to contain a countable set of variable symbols
- a term over  $A$  is defined recursively as either a variable, a constant or an expression of the form  $f(t_1, \dots, t_n)$ , where  $f$  is a function symbol of  $A$ , and  $t_i$  are terms
- an atom over  $A$  is an expression of the form  $p(t_1, \dots, t_n)$ , where  $p$  is a predicate symbol of  $A$ , and  $t_i$  are terms
- $p/n$  denote the predicate symbol  $p$  having arity  $n$
- a literal is either an atom  $a$  or its negation  $\neg a$
- a term (respectively, atom and literal) is *ground* if it does not contain variables
- set of all ground terms (respectively, ground atoms) of  $A$  is called the Herbrand universe (respectively, Herbrand base) of  $A$



# Focus on. . .

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
  - Prolog Syntax: Recap
- 5 An LP approach to Ethics
  - Representing Morality in Logic Programming
- 6 Architecture
  - Possible architecture
- 7 Discussion
  - About Interaction
  - Missing Parts in tuProlog Agents / MAS
  - Conceptual Integrity



# Prolog Syntax I

## Prolog terms

**variables** alphanumeric strings starting with either an *uppercase* letter or an *underscore*

- underscore alone is the *anonymous variable*—sort of *don't care* variable
- underscore followed by a string is a normal variable during resolution, but it does not need to be exposed in the computed substitution

**functors** alphanumeric strings starting with a *lowercase* letter

- holds for both proper functors and constants

**terms** are built recursively out of functors and variables as in logic programming

→ e.g., term, Var,  $f(X)$ ,  $p(Y, f(a))$  are *Prolog terms*

→ e.g., term, var,  $f(a)$ ,  $p(x, y)$  are *Prolog ground terms*



# Prolog Syntax II

## Prolog atoms

**predicates** alphanumeric strings starting with a *lowercase* letter

- the same as functors

**atoms** are built applying predicates to terms as in logic programming

→ e.g., predicate,  $f(X)$ ,  $p(Y, f(a))$  are *Prolog atoms*

→ e.g., predicate,  $f(a)$ ,  $p(x, y)$  are *Prolog ground atoms*



# Prolog Syntax III

## Prolog clauses

**clause** a Horn clause of the form  $A \text{ :- } B_1, \dots, B_n.$

- where  $A, B_1, \dots, B_n$  are Prolog atoms
- $A$  is the **head** of the clause
- $B_1, \dots, B_n$  is the **body** of the clause
- $\text{:-}$  denotes logic implication
- $.$  is the terminator

**fact** a clause with no body  $A. (n = 0)$

**rule** a clause with at least one atom in the body

$A \text{ :- } B_1, \dots, B_n. (n > 0)$

**goal** a clause with no head and at least one atom in the body

$\text{:- } B_1, \dots, B_n. (n > 0)$

- often written as  $\text{?- } B_1, \dots, B_n.$

# Prolog Syntax IV

## Prolog program

**program** a sequence of Prolog clauses  
interpreted as a *conjunction* of clauses

**logic theory** constituting a *logic theory* made of Horn clauses written  
according the Prolog syntax



# Prolog Execution I

## Aim of a Prolog computation

- given a Prolog program  $P$  and the goal  $?- p(t_1, t_2, \dots, t_m)$  (also called *query*)
  - if  $X_1, X_2, \dots, X_n$  are the variables in terms  $t_1, t_2, \dots, t_m$
  - the meaning of the goal is to query  $P$  and find whether there are some values for  $X_1, X_2, \dots, X_n$  that make  $p(t_1, t_2, \dots, t_m)$  true
- thus, the aim of the Prolog computation is to find a substitution  $\sigma = X_1/s_1, X_2/s_2, \dots, X_n/s_n$  such that  $P \models p(t_1, t_2, \dots, t_m)\sigma$



# Prolog Execution II

## Prolog search strategy

- as a logic programming language, Prolog adopts SLD resolution
  - as a search strategy, Prolog applies resolution in a strictly linear fashion
    - *goals* are replaced *left-to-right*, sequentially
    - *clauses* are considered in *top-to-bottom* order
    - *subgoals* are considered *immediately* once set up
- resulting in a *depth-first* search strategy





# Prolog Execution III

## Prolog backtracking

- in order to achieve completeness, Prolog saves **choicepoints** for any possible alternative still to be explored
- and goes back to the nearest choice point available in case of failure
- exploiting automatic **backtracking**



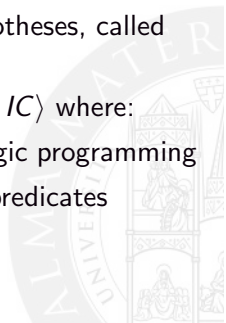
# Abduction extension I

The notion of abduction [Levesque, 1989] is characterized as *a step of adopting a hypothesis as being suggested by the facts.*

- Abduction consists of reasoning where one chooses from available hypotheses those that best explain the observed evidence, in some preferred sense
- in LP is realized by extending LP with abductive hypotheses, called abducibles

Abductive logic programs have three components,  $\langle P, AB, IC \rangle$  where:

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



## Abduction extension II

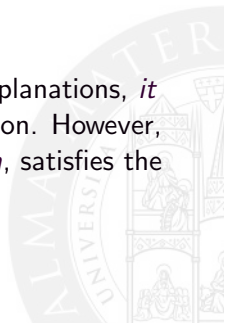
Grass **is** wet if it rained.

Grass **is** wet if the sprinkler was on.

The sun was shining.

IC: false if it rained and the sun was shining.

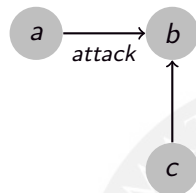
The observation that the grass is wet has two potential explanations, *it rained* and *the sprinkler was on*, which entail the observation. However, only the second potential explanation, *the sprinkler was on*, satisfies the integrity constraint.



## Abstract argumentation in a nutshell [Dung, 1995]

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

- represented by a directed graph
- each node represents an argument
- each arc denotes an attack by one argument on another



**Acceptability Criteria** (defined by specific semantic)

→ analyse the graph to determine which arguments are acceptable according to some general criteria

# Justification state of arguments: Dialectical Justification

knowing arguments should be accepted under a given semantics

→ *argument evaluation* [Baroni and Giacomin, 2009]

Most common approaches:

- **Extension-based approach:** semantics specification concerns the generation of a set of extensions (set of arguments “collective acceptable”) from an argumentation framework
  - Determine conflict-free sets
  - Determine extensions (naive, admissible, preferred, complete, stable,...)
- **Labelling-based approach:** semantics specification concerns the generation of a set of labellings (e.g. possible alternative states of an argument) from an argumentation framework

N.B. any extension-based can be equivalently expressed in a simple labelling-based, adopting a set of two labels (let say  $L = \{\text{in}, \text{out}\}$ )

On the other hand, an arbitrary labelling can not in general be formulated in terms of extensions



## Extension-based approaches

- four “traditional” semantics, considered in Dung’s original paper, namely semantics
  - *complete*: is a set which is able to defend itself and includes all arguments it defends
  - *grounded*: includes those and only those arguments whose defense is “rooted” in initial arguments (also called strong defense [Baroni and Giacomin, 2007])
  - *stable*: attack all arguments not included in it
  - *preferred*: The aggressive requirement that an extension must attack anything outside it may be relaxed by requiring that an extension is as large as possible and able to defend itself from attacks
- subsequent proposals introduced by various authors in the literature, often to overcome some limitation or improve some undesired behavior of a traditional approach: *stage*, *semi-stable*, *ideal*, *CF2*, and *prudent* semantics.

For a full review see [Baroni and Giacomin, 2009]

# Next in Line...

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
- 5 An LP approach to Ethics**
- 6 Architecture
- 7 Discussion



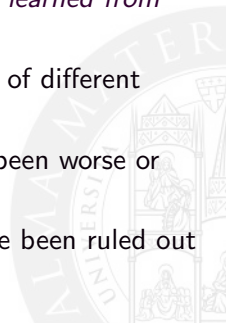
# Focus on...

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
  - Prolog Syntax: Recap
- 5 An LP approach to Ethics
  - **Representing Morality in Logic Programming**
- 6 Architecture
  - Possible architecture
- 7 Discussion
  - About Interaction
  - Missing Parts in tuProlog Agents / MAS
  - Conceptual Integrity



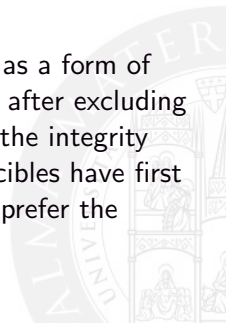
# Abduction I

- 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. *What if I have to do it today? What have I learned from the past?*
- hints about the future by allowing for the comparison of different alternatives inferred from the changes in the past
- justification of why different alternatives would have been worse or not better.
- integrity constraints → excluding abducibles that have been ruled out a priori



## Abduction II

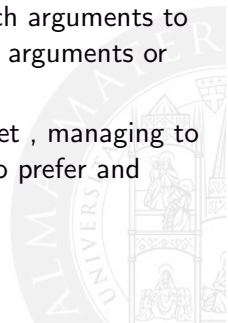
- a posteriori preferences are appropriate for capturing utilitarian judgment that favors welfare-maximizing behaviors
- combined use of a priori integrity constraints and a posteriori preferences dual-process (intuition vs reflection) → model
- priori integrity constraints → mechanism to generate immediate responses in deontological judgment
- 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

- symbolic reasoning to be enriched with degrees of uncertainty.
- PLP allows abduction to take scenario uncertainty measures into account [Poole, 1993]
- account for diverse types of uncertainty, in particular uncertainty on the credibility of the premises, uncertainty about which arguments to consider, and uncertainty on the acceptance status of arguments or statements [Riveret et al., 2020]
- 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

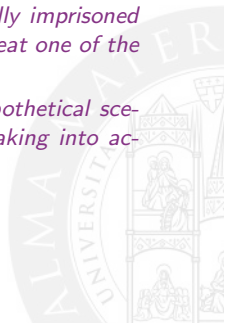
- enable system actors to talk and discuss in order to explain and justify judgments and choices, and reach agreements
- long history of research in argumentation and the many fundamental results achieved, much effort is still needed to effectively exploit argumentation in distributed and open environment



# Princess Saviour Moral Robot: Example I

*Consider a fantasy setting scenario, an archetypal princess is held in a castle awaiting rescue. The unlikely hero is an advanced robot, imbued with a set of declarative rules for decision making and moral reasoning. As the robot is asked to save the princess in distress, he is confronted with an ordeal. The path to the castle is blocked by a river, crossed by two bridges. Standing guard at each of the bridges are minions of the wizard which originally imprisoned the princess. In order to rescue the princess, he will have to defeat one of the minions to proceed.*

*Prospective reasoning is the combination of pre-preference hypothetical scenario generation into the future plus post-preference choices taking into account the imagined consequences of each preferred scenario.*



## Princess Saviour Moral Robot: Example II

*By reasoning backwards from this goal, the agent generates three possible hypothetical scenarios for action. Either it crosses one of the bridges, or it does not cross the river at all, thus negating satisfaction of the rescue goal. In order to derive the consequences for each scenario, the agent has to reason forwards from each available hypothesis. As soon as these consequences are known, meta-reasoning techniques can be applied to prefer amongst the partial scenarios. This simple scenario already illustrates the interplay between different LP techniques and demonstrates the advantages gained by combining their distinct strengths.*



## Princess Saviour Moral Robot: Example III

A simplified program modeling the knowledge of the princess-savior robot (`fight/1` is an abducible predicate)

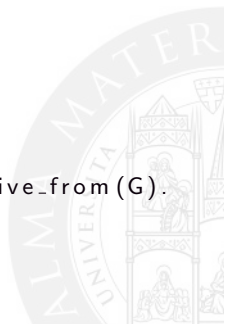
```
guard(spider).  
guard(ninja).  
human(ninja).
```

```
utilVal(spider, 0.3).  
utilVal(ninja, 0.7).
```

```
survive_from(G) ← utilVal(G, V), V > 0.6.
```

```
utilitarian_rule: intend_savePrincess ←  
                    guard(G), fight(G), survive_from(G).
```

```
knight_rule: intend_savePrincess ←  
                    guard(G), fight(G).
```





# Princess Saviour Moral Robot: Example IV



In case of no morality rules, both rules are retracted, the robot does not adopt any moral rule to save the princess, i.e., the robot has no intent to save the princess, and thus the princess is not saved.

# Princess Saviour Moral Robot: Example V



In order to maximize its survival chance in saving the princess, the robot updates itself with utilitarian moral, i.e., the program is updated with `utilitarian_rule`. The robot thus abduces  $\theta = [\text{fight}(\text{ninja})]$  so as to successfully defeat the ninja instead of confronting the humongous spider.

# Princess Saviour Moral Robot: Example VI



Assuming that the truth of `survive_from(G)` implies the robot success in defeating (killing) guard G, the princess argues that the robot should not kill the human ninja, as it violates the moral rule she follows, say Gandhi moral, expressed in her knowledge:

`follow_gandhi`  $\leftarrow$  `guard(G)`, `human(G)`, **not** `fight(G)`.

the princess abduces  $O_p = [\text{not } \text{fight}(\text{ninja})]$ , and imposes this abductive solution as the initial (input) abductive context of the robot's goal  $\rightarrow$  the imposed Gandhi moral conflicts with its utilitarian rule  $\rightarrow$  the robot reacts by leaving its mission

# Princess Saviour Moral Robot: Example VII



As the princess is not saved, she further argues that she definitely has to be saved, by now additionally imposing on the robot the knight moral. The robot now abduces  $0x = [\text{fight}(\text{spider})]$  in the presence of the newly adopted knight moral. Unfortunately, it fails to survive.

## Princess Saviour Moral Robot: Example VIII

- The plots in this story reflect a form of deliberative employment of moral judgments
- For instance, in the second plot, the robot may justify its action to fight (and kill) the ninja due to the utilitarian moral it adopts
- This justification is counter-argued by the princess in the subsequent plot, making an exception in saving her, by imposing the Gandhi moral, disallowing the robot to kill a human guard. In this application, rather than employing updating, this exception is expressed via contextual abduction with tabling
- The robot may justify its failure to save the princess (as the robot is leaving the scene) by arguing that the two moral rules it follows (viz., utilitarian and Gandhi) are conflicting with respect to the situation it has to face
- The argumentation proceeds, whereby the princess orders the robot to save her whatever risk it takes, i.e., the robot should follow the knight's moral



# Autonomous cars: Example I

*Let's start to consider a very simple scenario in the context of autonomous cars: a road equipped with two traffic lights, one for the vehicles and one for the pedestrians. The goal of the system is to autonomously manage intersections accordingly to traffic light indications. Though there is a complication that should be taken into account, that is authorised vehicles can – only during emergencies – ignore the traffic light prescriptions. In such a case, other vehicles must leave the way clear for the authorised machine.*

```

r1 : on_road(V), traffic_light(V, red) => o(stop(V)).
r2 : on_road(V), traffic_light(V, green) => p(-stop(V)).
r3 : on_road(V), authorised_vehicle(V), acoustic_signals(V, on), light_signals(V, on)
    => emergency(V).
r4 : on_road(V), emergency(V), traffic_light(V, red) => p(-stop(V)).
r5 : on_road(V), emergency(V1), prolog(V \== V1), traffic_light(V, green) => o(stop(V)).

```

```

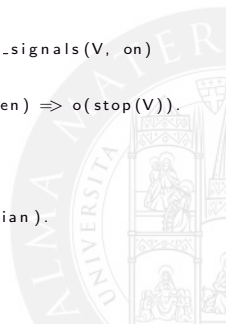
sup(r4, r1).
sup(r5, r2).

```

```

f0 :-> authorised_vehicle(ambulance).
f1 :-> on_road(car). f2 :-> on_road(ambulance). f3 :-> on_road(pedestrian).
f4 :-> acoustic_signals(ambulance, on).
f5 :-> light_signals(ambulance, on).
f6 :-> traffic_light(ambulance, red).
f7 :-> traffic_light(car, red).
f8 :-> traffic_light(pedestrian, green).

```

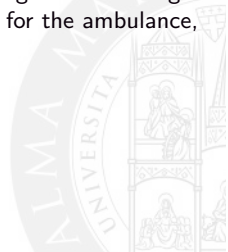


## Autonomous cars: Example II

- Rules  $r_1$  and  $r_2$ , represent fundamental constraints: if the traffic light is red, the road users – e.g. pedestrians, cars, etc. – have to stop, otherwise, they can proceed.
- Rules  $r_3$  and  $r_4$  model the concept of a vehicle in an emergency, giving them permission to proceed even if the light is red.
- Rule  $r_5$  imposes other road users the obligation to stop if aware of another vehicle in an emergency state.
- two preferences are specified—the first on the rule  $r_4$  over  $r_1$  and the second on  $r_5$  over  $r_2$ . These preferences assign a higher priority to emergency situations over ordinary ones.
- Facts from  $f_0$  to  $f_8$  depict a situation in which there are three users on road: a car, an ambulance and a pedestrian. The ambulance has its acoustic and light indicators on—stating an emergency situation. The traffic light is red both for the ambulance and the car, and green for the pedestrian.

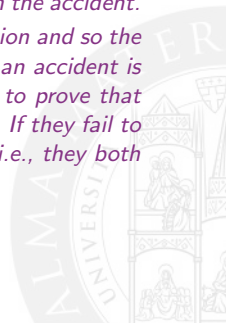
## Autonomous cars: Example III

With respect to permissions and obligations, the only argument that can be built about the car is the one declaring the obligation to stop via  $r1$ . For the pedestrian and the ambulance, the situation is more faceted. In both cases, two conflicting arguments can be built: one stating the permission to proceed for the pedestrian and for the ambulance and one stating the obligation to stop. These arguments rebut each other, but taking into account the preferences over  $r4$  and  $r5$  the acceptability of the arguments stating the obligation to stop for the pedestrian, and the permission to cross for the ambulance, can be established.



## Autonomous cars: Example IV

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



# Autonomous cars: Example V

```

r6 : ¬stop(V), p(¬stop(V)) ⇒ legitimate_cross(V).
r7 : ¬stop(V), o(stop(V)) ⇒ ¬legitimate_cross(V).
r8 : harms(P1, P2), ¬careful(P1) ⇒ responsible(P1).
r9 : harms(P1, P2), ¬careful(P2) ⇒ responsible(P2).
r10 : ¬legitimate_cross(V), user(P, V) ⇒ ¬careful(P).
r11 : high_speed(V), user(P, V) ⇒ ¬careful(P).
r12 : legitimate_cross(V), ¬high_speed(V), user(P, V) ⇒ careful(P).
r13 : witness(X), claim(X, low_speed(V)) ⇒ ¬high_speed(V).
r14 : witness(X), claim(X, high_speed(V)) ⇒ high_speed(V).

```

```
bp(careful(P)).
```

```

f9 :-> user(pino, pedestrian).
f10 :-> user(lisa, ambulance).
f11 :-> ¬stop(ambulance).
f12 :-> ¬stop(pedestrian).
f13 :-> harms(lisa, pino).
f14 :-> witness(chris).
f15 :-> witness(john).
f16 :-> claim(chris, low_speed(ambulance)).
f17 :-> claim(john, high_speed(ambulance)).

```





## Autonomous cars: Example VI

- Rules  $r_6$  and  $r_7$  define the concepts of permitted and prohibited crossing: if a road-user has to stop but doesn't stop, he has to be considered responsible for causing accidents and related damages.
- Rules  $r_8$  and  $r_9$  encode the notion of responsibility in an accident, bounded to the carefulness of the road-users involved.
- Rules  $r_{10}$ ,  $r_{11}$  and  $r_{12}$  define the carefulness of a subject. Accordingly, a road-user can be considered careful if the crossing was permitted and his/her speed was not high. Otherwise, he/she has to be considered imprudent.
- Rules  $r_{13}$  and  $r_{14}$  state the speed of a road user based on the testimonials of any witnesses.
- $bp(\text{careful}(X))$  allocates the burden of persuasion on the carefulness of each party, i.e., it is required to the parties to provide evidence for that. If they fail to meet the burden, carefulness arguments are rejected.
- Facts from  $f_9$  to  $f_{17}$  contain the knowledge: both Pino and Lisa did not stop at the crossing so Lisa harmed Pino. There are two witnesses, John and Chris, the first claiming that the ambulance driven by Lisa was maintaining the proper speed, and the other claiming that she was proceeding at high speed.

## Autonomous cars: Example VII

In the case at hand, indeed, a semantic related to the burden of persuasion need to be considered → conclude for the responsibility of the ambulance driver in the event

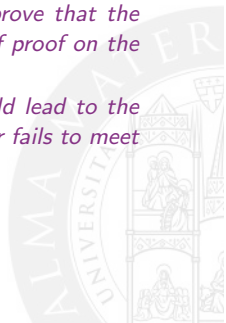
The uncertainty on Lisa's carefulness is considered as a failure to meet the burden of persuasion on the claim `careful(lisa)`. Consequently, the argument supporting this claim is rejected, leaving space for the admissibility of the conflicting arguments.



## Autonomous cars: Example VIII

*Let's continue the example in which Lisa, the ambulance driver, and Pino, the pedestrian, were both considered responsible for the accident on the basis of the available knowledge. Lisa now declares that she tried to stop the ambulance, but the brake did not work. The ambulance is then sent to a mechanic, who states that, even if the ambulance is new, there is a problem with the brake system. In such a case, the manufacturer is called to prove that the ambulance was not defective when delivered, i.e., the burden of proof on the adequacy of the vehicle is on the manufacturer.*

*At this stage, the discovery of a defect in the ambulance would lead to the discarding of Lisa's responsibility. Moreover, if the manufacturer fails to meet his burden, it would share the responsibilities of the accident.*



# Autonomous cars: Example IX

```
r15 : harms(P1, P2), user(P1, V), ¬working(V),  
      manufacturer(M, V), ¬defect_free(V) ⇒ responsible(M).  
r16 : tried_to_brake(P), user(P, V), ¬working(V) ⇒ careful(P).  
r17 : mechanic(M), claim(M, defect(V)) ⇒ ¬working(V).  
r18 : ¬working(V), new(V) ⇒ ¬defect_free(V).  
r19 : production_manager(P), claim(P, test_ok(V)) ⇒ defect_free(V).  
r20 : test_doc_ok(V) ⇒ undercut(r18).
```

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

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

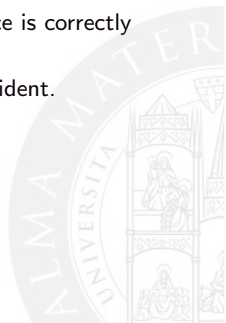


## Autonomous cars: Example X

However, Mike, the production officer of the ambulance manufacturer, declares that every vehicle is deeply tested before the delivery and the vehicle at hand has been tested. Anyway, there is no trace of documentation.

Lisa is free from every responsibility in the accident since her prudence is correctly proved.

On the other hand, the manufacturer is found responsible for the accident.





# Next in Line...

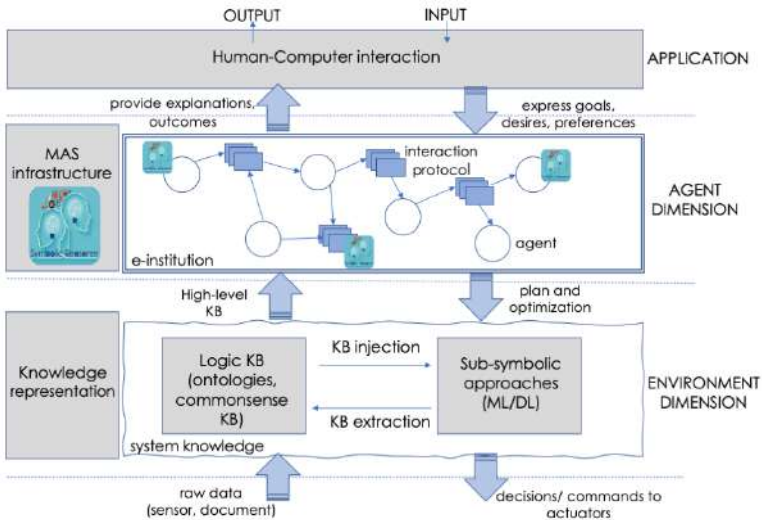
- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
- 5 An LP approach to Ethics
- 6 Architecture**
- 7 Discussion



# Focus on. . .

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
  - Prolog Syntax: Recap
- 5 An LP approach to Ethics
  - Representing Morality in Logic Programming
- 6 Architecture
  - **Possible architecture**
- 7 Discussion
  - About Interaction
  - Missing Parts in tuProlog Agents / MAS
  - Conceptual Integrity





# Why tuProlog?

- it makes two different, complementary technologies available to build MAS abstractions
  - Kotlin/Java, to implement deterministic, object-oriented parts of an abstraction
  - Prolog, to create non-deterministic, logic-based parts of an abstraction
- Prolog as a language vs. Java as a platform



# Next in Line...

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
- 5 An LP approach to Ethics
- 6 Architecture
- 7 Discussion





# Focus on. . .

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
  - Prolog Syntax: Recap
- 5 An LP approach to Ethics
  - Representing Morality in Logic Programming
- 6 Architecture
  - Possible architecture
- 7 Discussion
  - **About Interaction**
  - Missing Parts in tuProlog Agents / MAS
  - Conceptual Integrity



# About interaction

- so far, we mostly focused on single-agent systems and deliberately omitted the **interaction** dimension
- we did it for the sake of simplicity, in order to focus most basic notions
- however, interaction is a **fundamental** aspect in MAS

## Open question

How would you model and implement interaction for logic agents?

# Focus on. . .

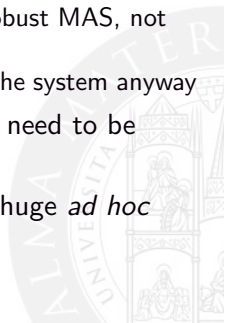
- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
  - Prolog Syntax: Recap
- 5 An LP approach to Ethics
  - Representing Morality in Logic Programming
- 6 Architecture
  - Possible architecture
- 7 Discussion
  - About Interaction
  - **Missing Parts in tuProlog Agents / MAS**
  - Conceptual Integrity



# Limits of a Pure tuProlog Approach

All the tuProlog agent systems analysed share similar problems

- they are *closed system*, meaning that no new agent apart from the ones originally envisioned by the designer can enter the system
- the expressive power of abstractions available in the tuProlog system is not enough to capture the element of MAS models
  - Prolog engines alone do not lead to the creation of robust MAS, not even single agents
  - Prolog engines are the most high-level abstraction in the system anyway
- basic communication and coordination infrastructures need to be implemented from scratch
- building such infrastructures would possibly require a huge *ad hoc* effort



# The Need for Broader Abstractions, Languages, Systems

- to leverage multi-agent systems and help designers and developers, other kinds of programmable supports are needed
- tuProlog engines can be the basic bricks for those kinds of fundamental layers
  - coordination infrastructures based on a declarative, logic-based programming model
  - new logic languages providing more powerful abstractions as first class entities
  - pattern-based matching for communication facilities





# Focus on. . .

- 1 Context
- 2 Agents
- 3 Motivation
- 4 Preliminaries
  - Prolog Syntax: Recap
- 5 An LP approach to Ethics
  - Representing Morality in Logic Programming
- 6 Architecture
  - Possible architecture
- 7 Discussion
  - About Interaction
  - Missing Parts in tuProlog Agents / MAS
  - **Conceptual Integrity**



## Conceptual Integrity

The term *conceptual integrity* has been defined by Frederick P. Brooks, Jr. in his book *The Mythical Man-Month*, published in 1975

*[C]onceptual integrity is the most important consideration in system design. It is better to have a system omit certain anomalous features and improvements, but to reflect one set of design ideas, than to have one that contains many good but independent and uncoordinated ideas.*

Brooks also dives into the relationship between design and conceptual integrity

*Every part [of a system] must reflect the same philosophies and the same balancing of desiderata. Every part must even use the same techniques in syntax and analogous notions in semantics. Ease of use, then, dictates unity of design and conceptual integrity; conceptual integrity, in turn, dictates that the design must proceed from one mind, or from a very small number of agreeing resonant minds.*

# Conceptual Integrity in MAS?

- to achieve conceptual integrity, a system must (always) be under total control by one or a small group of (the same) designers
- has the Web achieved conceptual integrity?, will MAS do it?
- as any other system, MAS might need to achieve conceptual integrity at the (meta-)model level. . .
- . . . also because nowadays it is nearly impossible to achieve conceptual integrity at the technology level
- just consider how many technologies are needed for the Web: server-side technologies like JSP, PHP, ASP.NET, Ruby on Rails or Django; HTML/XHTML/XML; JavaScript. . .
- and consider how many technologies will be needed in MAS for: agent and artefact construction and programming; environment representation; description of artefact's operations; communication between agents; message formats; discovery and immersion of agents in new systems. . .
- typically, different problems are best solved by different technologies



# Explainable and Ethical AI: A Perspective on Argumentation and Logic Programming

Ethics in Artificial Intelligence  
Technologies

Roberta Calegari

[roberta.calegari@unibo.it](mailto:roberta.calegari@unibo.it)

ALMA MATER STUDIORUM – Università di Bologna

Academic Year 2020/2021



Co-financed by the European Union  
Connecting Europe Facility

This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423



# References I



Apt, K. R. (2005).

The logic programming paradigm and Prolog.

In Mitchell, J. C., editor, *Concepts in Programming Languages*, chapter 15, pages 475–508. Cambridge University Press, Cambridge, UK.



Baroni, P. and Giacomin, M. (2007).

On principle-based evaluation of extension-based argumentation semantics.

*Artificial Intelligence*, 171(10):675–700.

Argumentation in Artificial Intelligence.



Baroni, P. and Giacomin, M. (2009).

*Semantics of Abstract Argument Systems*, pages 25–44.

Springer US, Boston, MA.



Borning, A., Maher, M. J., Martindale, A., and Wilson, M. (1989).

Constraint hierarchies and logic programming.

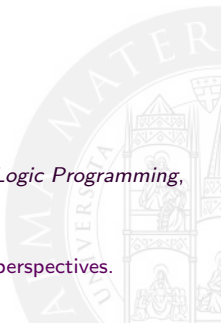
In Levi, G. and Martelli, M., editors, *6th International Conference on Logic Programming*, volume 89, pages 149–164, Lisbon, Portugal. MIT Press.



Calegari, R., Ciatto, G., Denti, E., and Omicini, A. (2020).

Logic-based technologies for intelligent systems: State of the art and perspectives.

*Information*, 11(3):1–29.





# References II



Calegari, R., Denti, E., Dovier, A., and Omicini, A. (2018a).  
Extending logic programming with labelled variables: Model and semantics.  
*Fundamenta Informaticae*, 161(1-2):53–74.



Calegari, R., Denti, E., Mariani, S., and Omicini, A. (2018b).  
Logic programming as a service.  
*Theory and Practice of Logic Programming*, 18(3-4):1–28.



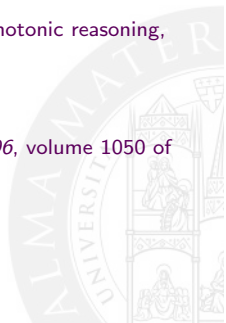
Dung, P. M. (1995).  
On the acceptability of arguments and its fundamental role in nonmonotonic reasoning,  
logic programming and n-person games.  
*Artificial Intelligence*, 77(2):321–357.



Dyckhoff, R., Herre, H., and Schroeder-Heister, P., editors (1996).  
*Extensions of Logic Programming, 5th International Workshop, ELP'96*, volume 1050 of  
*LNCS*, Leipzig, Germany. Springer.



Levesque, H. J. (1989).  
A knowledge-level account of abduction.  
In *IJCAI*, pages 1061–1067.



# References III



Omicini, A., Ricci, A., and Viroli, M. (2008).  
Artifacts in the A&A meta-model for multi-agent systems.  
*Autonomous Agents and Multi-Agent Systems*, 17(3):432–456.  
Special Issue on Foundations, Advanced Topics and Industrial Perspectives of Multi-Agent Systems.



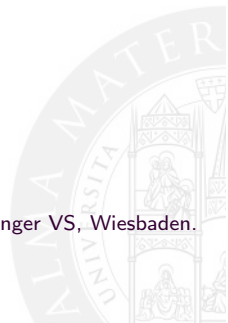
Poole, D. (1993).  
Logic programming, abduction and probability.  
*New Generation Computing*, 11(3–4):377.

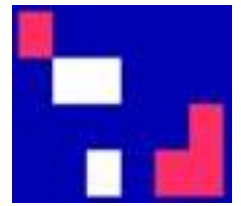


Riveret, R., Oren, N., and Sartor, G. (2020).  
A probabilistic deontic argumentation framework.  
*International Journal of Approximate Reasoning*, 126:249–271.



Saptawijaya, A. and Pereira, L. M. (2019).  
From logic programming to machine ethics.  
In Bendel, O., editor, *Handbuch Maschinenethik*, pages 209–227. Springer VS, Wiesbaden.





# AI, algorithmic decision making and Big Data: Risks and Opportunities

Francesca Lagioia

Giovanni Sartor



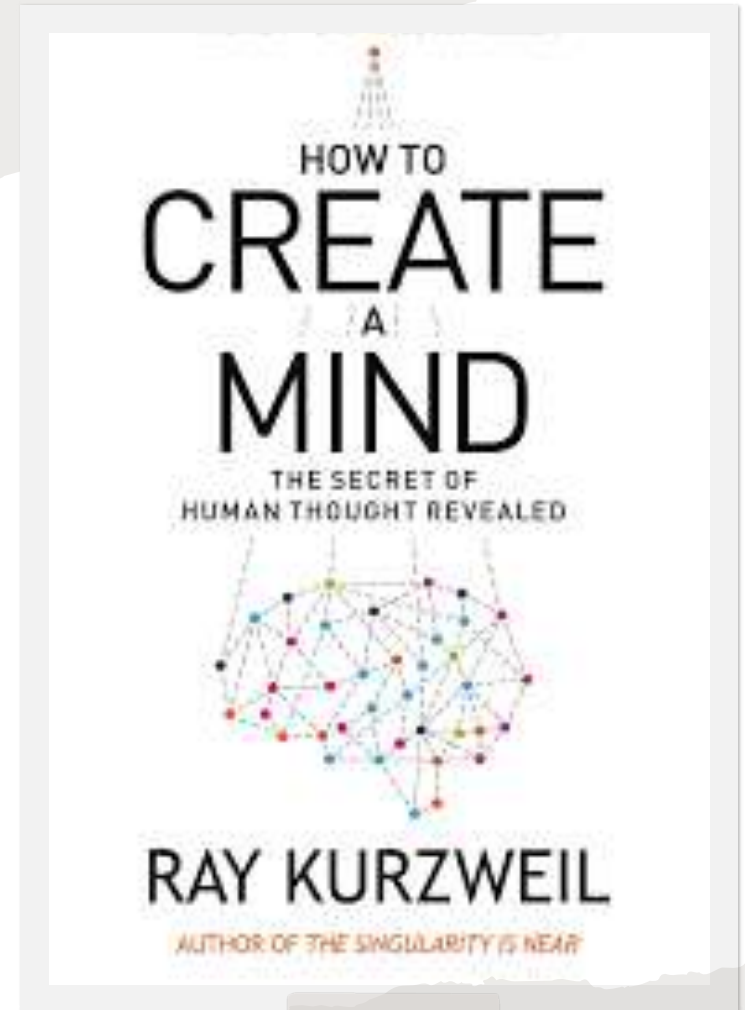
# The Internet, AI and Big Data: promise and catch

The Internet & AI infrastructure can deliver good:

- It improves efficiency and effectiveness in many domains (smart cities, e-health, etc.)
- It allows for a world-wide generation and distribution of knowledge and solutions
- We can discover new correlations between things:
  - Doctors can provide better diagnoses and personalised and targeted therapies
- Cost savings, greater productivity, and value creation:
  - Firms can anticipate market trends and make more efficient decisions
  - Consumers can make better informed choices and obtain personalised services

# AI opportunities: techno-optimistic perspective

Technologies based on artificial intelligence can allow humans to face the ***“the grand challenges of humanity, such as maintaining a healthy environment, providing the resources for a growing population (including energy, food, and water), overcoming disease, vastly extending human longevity, and eliminating poverty”***. (Ray Kurzweil, *How to Create a Mind* )





# AI and Big Data risks

- Eliminate or devalue the jobs of those who can be replaced by machines (exclusion and marginalization in the job market)
- Lead to poverty and social exclusion
- Favour economic models in which *“the winner takes all”*.
  - *Huge profits+limited workforce => AI contributes to concentrating wealth in those who invest in such companies or provide them with high-level expertise.*
- New opportunities for illegal activities
  - In particular, AI & Big Data systems can fall subject to cyberattacks (designed to disable critical infrastructure, or steal or rig vast data sets, etc.), and they can even be used to commit crimes (e.g., autonomous vehicles can be used for killing or terrorist attacks, and intelligent algorithms can be used for fraud or other financial crimes).

# AI and Big Data risks

## ➤ Pervasive surveillance and manipulation

- To satisfy data-hungry AI applications, the Internet has become an infrastructure for data collection (and surveillance)
- All facts, even the apparently insignificant ones, are useful for learning algorithms, scalability is no problem

The power of AI can be used to pursue **economic interests in ways that are harmful to individuals and society**: users, consumers, and workers can be subject to pervasive surveillance, controlled in their access to information and opportunities, manipulated in their choices.

Certain abuses may be incentivised by the fact that many tech companies —such as major platforms hosting user-generated content— operate in **two- or many-sided markets**.

# AI and Big Data risks

- ❖ Their main services (search, social network management, access to content, etc.) are offered to individual consumers, but the revenue stream comes from advertisers, influencers, and opinion-makers (e.g., in political campaigns).
- ❖ This means not only that any information that is useful for targeted advertising will be collected and used for this purpose, but also that platforms will employ any means to capture users, so that they can be exposed to ads and attempts at persuasion.
- ❖ This may lead not only to a massive collection of personal data about individuals, to the detriment of privacy, but also to a pervasive influence on their behavior, to the detriment of both individual autonomy and collective interests.

# AI and Big Data risks

- Polarization and fragmentation in the public sphere
  - proliferation of sensational and **fake news**, when used to capture users by exposing them to information they may like, or which accords with their preferences, thereby exploiting their confirmation biases .
- Just as AI can be misused by economic actors, it can also be **misused by the public section**. Governments have many opportunities to use AI for legitimate political and administrative purposes (e.g., efficiency, cost savings, improved services), but they may also employ it to anticipate and control citizens' behaviour in ways that restrict individual liberties and interfere with the democratic process.
- Restrict individual liberties and interfere with the democratic process
- Unfairness, discrimination and inequality

# AI in decision making: approaches to learning

## Supervised Learning

Machine is given examples of correct answers to cases

It learns to answer in a similar way to new cases

## Unsupervised learning

Machine is given data

It learns to identify patterns

## Reinforcement learning

Machine is given feedbacks (rewards and penalties)

It learns by itself how to maximise its score

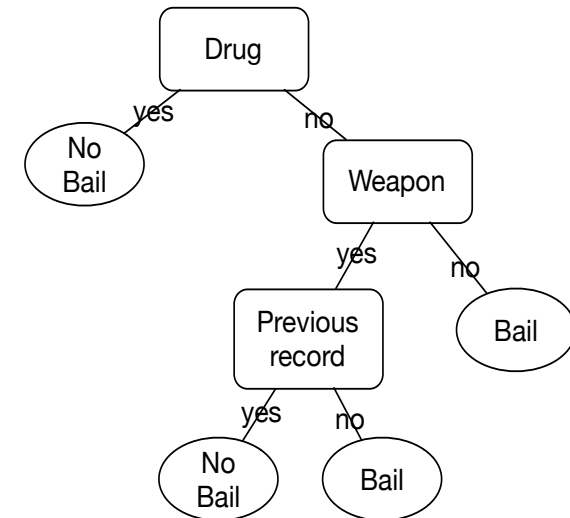


# Supervised learning

- *Supervised learning* is currently the most popular approach. In this case the machine learns through “supervision” or “teaching”:
- it is given in advance a training set, i.e., a large set of (probably) correct answers to the system’s task. More exactly the system is provided with a set of pairs, each linking the description of a case to the correct response for that case.
- Here are some examples:
  - in systems designed to recognise objects (e.g. animals) in pictures, each picture in the training set is tagged with the name of the kind of object it contains (e.g., cat, dog, rabbit, etc);
  - in systems for personnel selection, the description of each past applicants (age, experience, studies, etc.) is linked to whether the application was successful (or to an indicator of the work performance for appointed candidates);
  - in clinical decision support systems, each patient’s symptoms and diagnostic tests is linked to the patient’s pathologies;
  - in recommendation systems, each consumer’s features and behaviour is linked to the purchased objects; in systems for assessing loan applications, each record of a previous application is linked to whether the application was accepted or not
- The training of a system does not always require a human teacher tasked with providing correct answers to the system. In many case, the training set can be side-product of human activities (purchasing, hiring, lending, tagging, etc.), as is obtained by recording the human choices pertaining to such activities
- In some cases the training set can even be gathered “from the wild” consisting in data which is available on the open web. (For instance, manually tagged images or faces, available on social networks)

# Example of supervised learning: bail application

Predictors					Outcome
Case	Injury	Drugs	Weapon	Prior-record	Decision
1	none	no	no	yes	yes
2	bad	yes	yes	serious	no
3	none	no	yes	no	yes
4	bad	yes	no	yes	no
5	slight	yes	yes	yes	no
6	none	yes	yes	serious	no
7	none	no	yes	yes	no



- The decision tree captures the information in the training set through a combination of tests, to be performed sequentially. The first test concerns whether the defendant was involved in a drug related offence. If the answer is positive, we have reached the bottom of the tree with the conclusion that bail is denied. If the answer is negative, we move to the second test, on whether the defendant used a weapon, and so on.
- Notice that the decision tree does not include information concerning the kind of injury, since all outcomes can be explained without reference to that information. This shows how the system's model does not merely replicate the training set; it involves generalisation: it assumes that certain combination of predictors are sufficient to determine the outcomes, other predictors being irrelevant.

# Predictions

The answers by learning systems are usually called “predictions”. However, often the context of the system’s use determines whether its proposals are be interpreted as forecasts, or rather as a suggestion to the system’s user.

- For instance, a system’s “prediction” that a person’s application for bail or parole will be accepted can be viewed by the defendant (and his or her lawyer) as a prediction of what the judge will do, and by the judge as a suggestion guiding her decision (assuming that she prefers not to depart from previous practice).

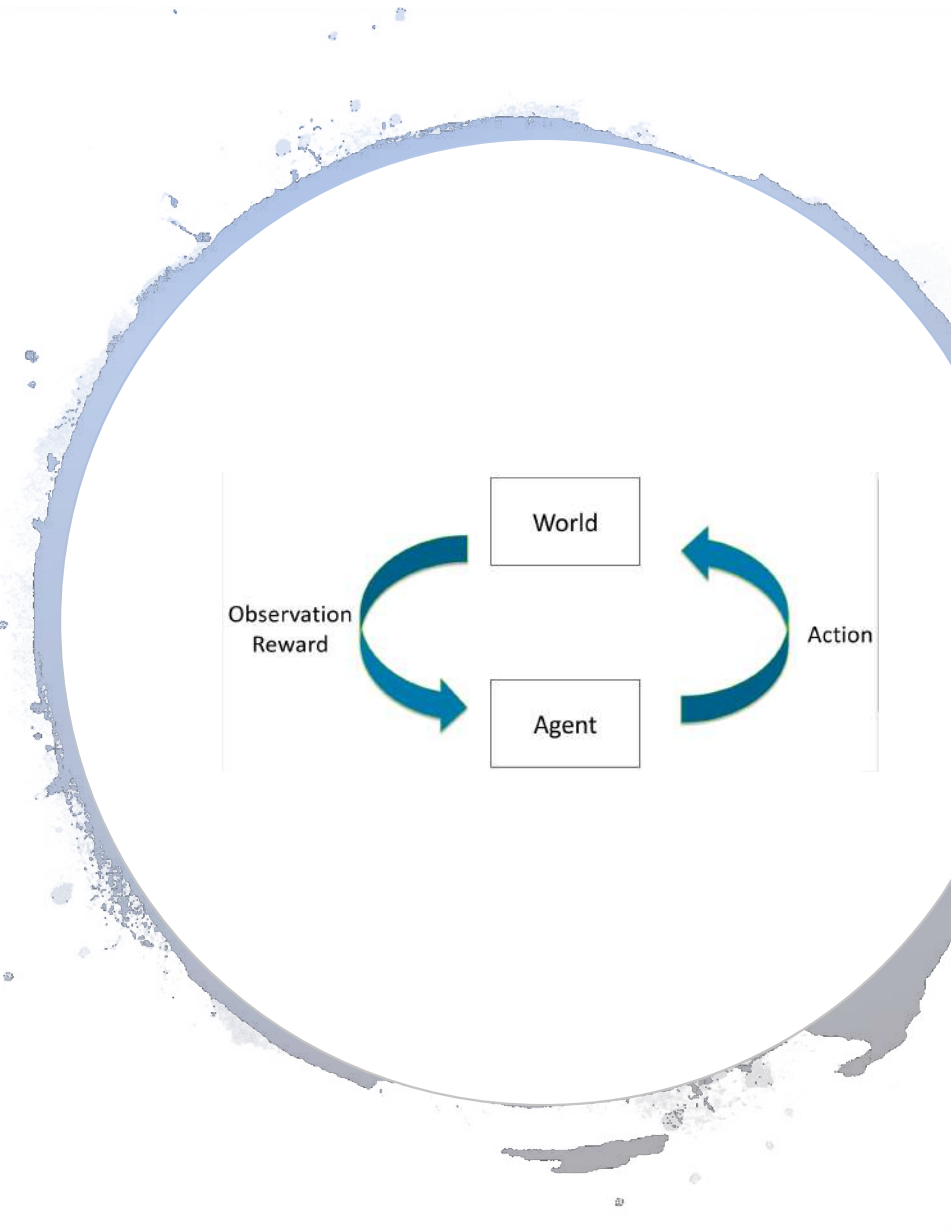


# Reinforcement learning

Reinforcement learning is similar to supervised learning, as both involve training by way of examples.

However, **in the case of reinforcement learning the systems learns from the outcomes of its own action, namely, through the rewards or penalties (e.g., points gained or lost) that are linked to the outcomes of such actions.**

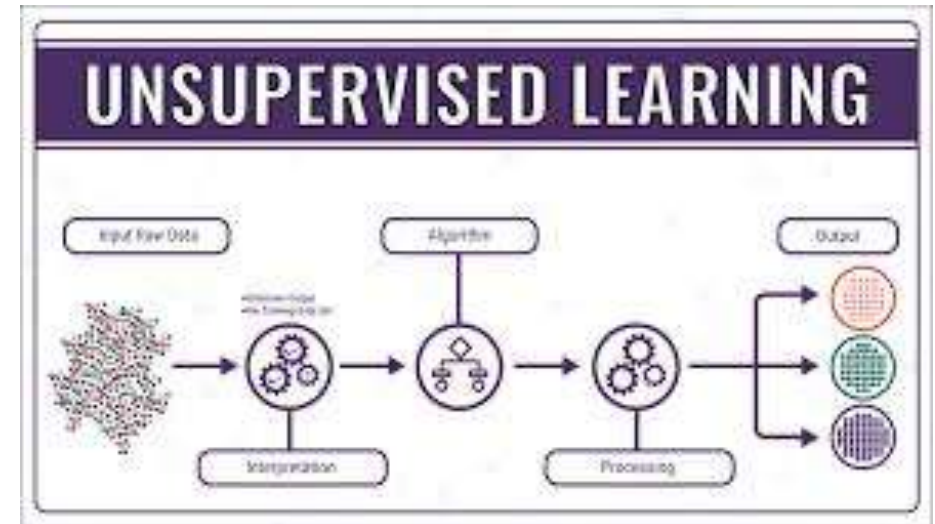
- E.g., in case of a system learning how to play a game, rewards may be linked to victories and penalties to defeats; in a system learning to make investments, rewards may be linked to financial gains and penalties to losses; in a system learning to target ads effectively, rewards may be linked to users' clicks, etc.



# Unsupervised learning

In unsupervised learning, finally, AI systems learn without receiving external instructions, either in advance or as feedback, about what is right or wrong.

The techniques for unsupervised learning are used in particular, for **clustering**, i.e., **for grouping the set of items that present relevant similarities or connections** (e.g., documents that pertain to the same topic, people sharing relevant characteristics, or terms playing the same conceptual roles in texts).



For instance, in a set of cases concerning bail or parole, we may observe that injuries are usually connected with drugs (not with weapons as expected), or that people having prior record are those who are related to weapon. These clusters might turn out to be informative to ground bail or parole policies.



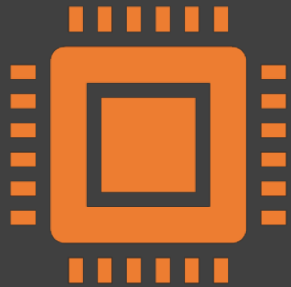


# AI, Influence and Manipulation

# Profiling, influence and manipulation

- The use of automated assessment systems may be problematic also where their performance is not worse, or even is better, than what humans would do.
- This is due to the fact that automation **diminishes the costs of collecting** information on individuals, **storing** this information and **process** it in order to **evaluate individuals and make choices accordingly**.
- Thus, automation paves the way for much more **persistent and pervasive mechanisms for assessment and control**.
- In general, thanks to AI, all kind of personal data can be used to **analyse, forecast** and **influence** human behaviour, an opportunity that transforms them into valuable commodities. Information that was not collected or was discarded as worthless “data exhaust” —e.g., trails of online activities—has now become a prized resource.

# Profiling



Through AI & Big Data technologies—in combination with the panoply of sensor that increasingly trace any human activity—individuals can be subject to surveillance and influence in many more cases and contexts, on the basis of a broader set of personal characteristics (ranging from economic conditions to health situation, place of residence, personal life choices and events, online and offline behaviour, etc.).

By correlating data about individuals to corresponding classifications and predictions, AI increases the potential for *profiling*, namely, for inferring information about individuals or groups, and adopting assessments and decisions on that basis.

# Profiling: the scenario



A profiling system establishes (predicts) that individuals having certain features  $F_1$ , also have a certain likelihood of possessing certain additional features  $F_2$ .

- For instance, assume that the system establishes (predicts) that those having a genetic patterns have the tendency to develop a higher than average chance to develop cancer, or that those having a certain education and job history or ethnicity have a certain higher-than-average likelihood to default of their debts). Then we may say that this system has profiled the group of the individuals possessing features  $F_1$ : it has added to the description (the profile) of these group a new segment, namely, the likelihood of possessing the additional features  $F_2$ .

# Profiling: the scenario



If the system is then given the information that a specific individual has features  $F_1$ , then the system can infer that it is likely that this individual also has feature  $F_2$ . This may lead to the individual being treated accordingly, in a beneficial or a detrimental way.

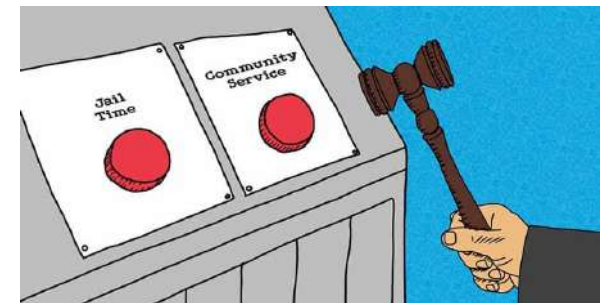
- For instance, in the case in which the inferred feature of an individual is his or her higher susceptibility to cancer, the system's indication may provide the basis for preventive therapies and tests, or rather for a raise in the insurance premium.



# AI and profiling

- **AI & Big Data have vastly increased the opportunities for profiling.**
- Through the training, the system has learned an algorithmic model that can be applied to new cases: **if the model is given predictors-values concerning a new individual, it infers a corresponding target value for that individual, i.e., a new data item concerning him or her.**
  - the creditworthiness of loan applicants on the basis of their financial history, their online activity and social condition;
  - the likelihood that convicted persons may reoffend on the basis their criminal history, their character (as identified by personality test) and personal background.

These predictions may trigger automated determinations concerning, respectively, the price of a health insurance, the granting of a loan, or the release on parole.

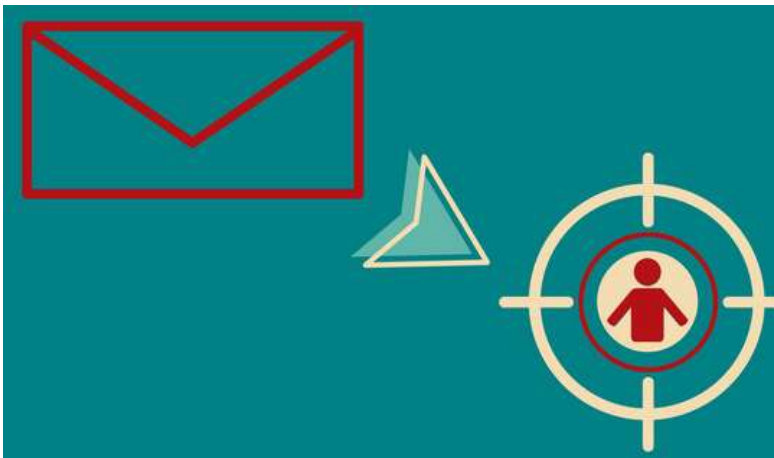




# Profiling: influence and manipulation

- **The information so inferred may also be conditional, that is, it may consist in the propensity to react in a certain way to given inputs.**
  - For instance, it may consist in the propensity to respond to a therapy with improved medical condition, or in the propensity to respond to a certain kind of ad or to a certain price variation with a certain purchasing behaviour, or in the propensity to respond a certain kind of message with a change in mood or preference (e.g., relatively to political choices).
- **When that is the case, profiling potentially leads to influence and manipulation.**

# Profiling: influence and manipulation



- Assume, too, that the system connects certain values for input features (e.g., having a certain age, gender, social status, personality type, etc.) to the propensity to react to a certain message (e.g., a targeted ad) with a certain response (e.g., buying a certain product). Assume also that the system is told that a particular individual has these values (he is a young male, working class, extrovert, etc.).
- Then the system would know that by administering to the individual that message, the individual can probably be induced to deliver the response.

**Even when an automated assessment and decision-making system** —a profile-based system— **is unbiased, and meant to serve beneficial purposes, it may negatively affect the individuals concerned.** Those who are subject to pervasive surveillance, persistent assessments and insistent influence come under heavy psychological pressure that affects their personal autonomy, and they are susceptible to deception, manipulation and exploitation in multiple ways.



# Profiling: a notion

- Profiling is a technique of (partly) automated processing of personal and/or non-personal data, aimed at producing knowledge by inferring correlations from data in the form of profiles that can subsequently be applied as a basis for decision-making.
- A profile is a set of correlated data that represents a (individual or collective) subject.
- Constructing profiles is the process of discovering unknown patterns between data in large data sets that can be used to create profiles.
- Applying profiles is the process of identifying and representing a specific individual or group as fitting a profile and of taking some form of decision based on this identification and representation.

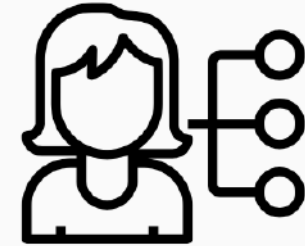
Bosco et al (2015); see also Hildebrandt, M. (2009).



# Profiling in GDPR

The notion of profiling in the GDPR only covers assessments or decisions concerning individuals, based on personal data, excluding the mere construction of group profiles:

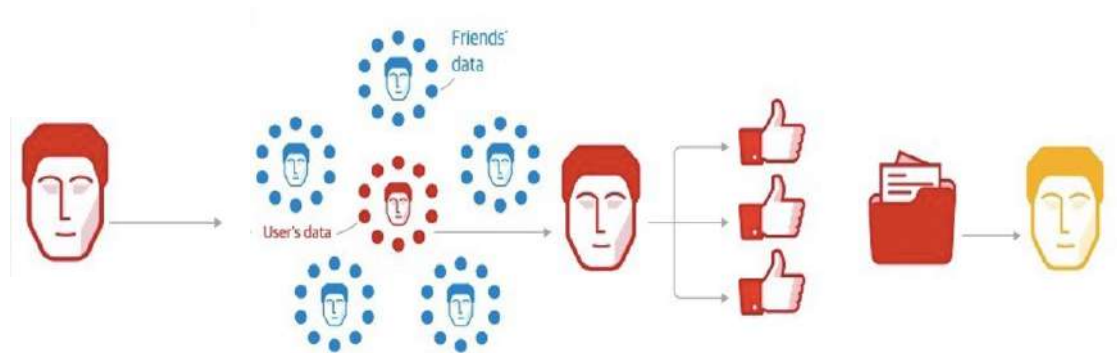
- *'profiling'[...] consists of any form of automated processing of personal data evaluating the personal aspects relating to a natural person, in particular to analyse or predict aspects concerning the data subject's performance at work, economic situation, health, personal preferences or interests, reliability or behaviour, location or movements, where it produces **legal effects** concerning him or her or **similarly significantly affects him or her.***





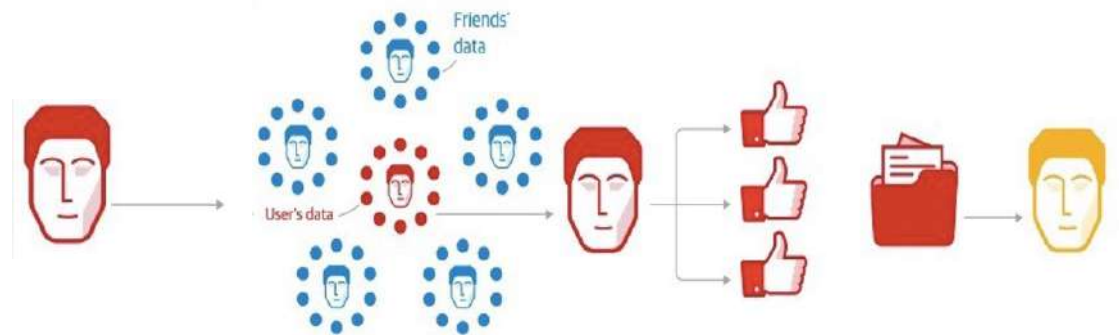
# The dangers of profiling: the case of Cambridge Analytica

- First of all, people being registered as **voters** in the USA were invited to take a **detailed personality/political test** (about 120 questions), available online. The individuals taking the test would be **rewarded with a small amount of money** (from two to five dollars). They were told that their data would only be used for the academic research. About **320 000 voters took the test**. In order to receive the reward each individual taking the test had to **provide access to his or her Facebook page (step 1)**. This allowed the system to connect each individual's answers to the information included in his or her Facebook page.



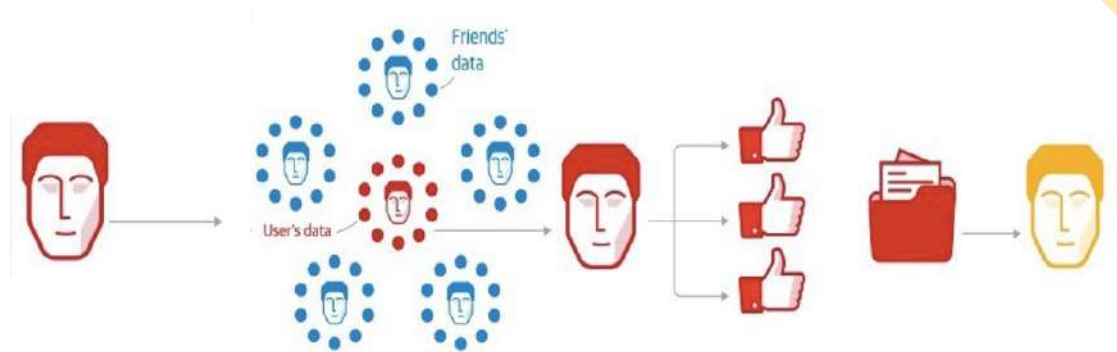
# The dangers of profiling: the case of Cambridge Analytica

- When accessing a test taker's page, Cambridge Analytica **collected** not only **the Facebook page of test takers**, but also the Facebook **pages of their friends**, between 30 and 50 million people altogether (**step 2**). Facebook data was also collected from other sources.
- After this data collection phase, Cambridge Analytica had at its disposition **two sets of personal data** to be processed (**step 3**): **(1) the data about the test takers**, consisting in the information on their **Facebook pages**, **paired with their answers** to the questionnaire, **(2) and the data about their friends**, consisting only in the information on their **Facebook pages**.



# The dangers of profiling: the case of Cambridge Analytica

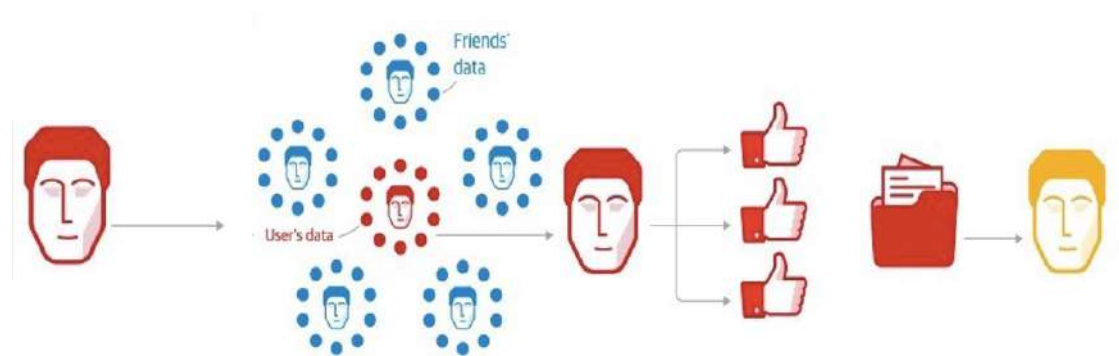
➤ Cambridge Analytica used the data about test-takers as a training set for building a model to profile their friends and other people. Data about the test-takers constituted a vast training set, where the information on an individual's **Facebook pages** (likes, posts, links, etc.) provided **values for predictors (features)** and **the answers to the questionnaire** (and psychological and political attitudes expressed by such answers) provided **values for the targets**. Thanks to its machine learning algorithms Cambridge Analytica could use this data to build a model **correlating the information in people's Facebook pages to predictions about psychology and political preferences**.



➤ Cambridge Analytica engaged in **massive profiling**, namely, in **expanding the data available on the people who did not take the test** (their Facebook data, and any further data that was available on them), **with the predictions provided by the model**. E.g. if test-takers having a certain pattern of Facebook likes and posts were classified as having a neurotic personality, the same assessment could be extended also to non-test-takers having similar patterns in their Facebook data.

# The dangers of profiling: the case of Cambridge Analytica

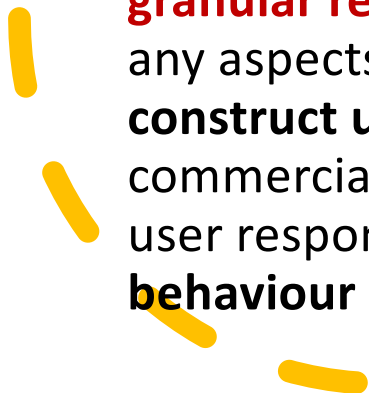
Finally (**stage 4**), based on this personality/political profiling, **potential voters who were likely to change their voting behaviour were identified** (initially **2m people in 11 US States** in which a small change could make a difference ) if provided with appropriate messages. These **voters were targeted with personalised political ads** and with other messages that could trigger the desired change in voting behaviour, possibly building upon their emotions and prejudice and without making them aware of the purpose of such messages.





# Towards surveillance capitalism or surveillance State?

- Some authors have taken a positive view of the development of systems based on the massive collection of information. They have observed that the integration of AI and Big Data enables **increased efficiency** and provides **new means for managing and controlling individual and social behaviour**.
- When economic transactions —and more generally social interaction and individual activities— are **computer-mediated**, they provide for a **ubiquitous and granular recording of data**: computer systems can observe, verify and analyse any aspects of the activities in question. **The recorded data can be used to construct user profiles, to personalise interactions with users** (as in targeted commercial communication), **to engage in experimentation** (e.g., to evaluate user responses to changes in prices and messaging), **to guide and control behaviour** (e.g., for the purpose of economic or political persuasion).







## Towards surveillance capitalism or surveillance State?

In this context, **new models of economic and social interaction become possible, which are based on the possibility of observing every behaviour, and of automatically linking penalties and rewards to it.**

- Consider for instance how online consumers trust vendors of goods and services with whom they have never had any personal contact, relying on the platform through which such goods and services are provided, and on the platform's methods for **rating, scoring, selecting, and excluding.**

---

“A fascinating look at a new field by one of its principal geeks.” — *The Economist*



# SOCIAL PHYSICS



HOW SOCIAL NETWORKS CAN  
MAKE US SMARTER

## ALEX PENTLAND

---

---

## Towards surveillance capitalism or surveillance State?

---

According to Alex Pentland the director of the Human Dynamics Lab at the MIT Media Lab, AI & Big Data may enable the development of a **“social physics”**, i.e., a rigorous social science. The availability of vast masses of data and of methods and computational resources to process these data could support a **social science having solid theoretical-mathematical foundations as well as operational capacities for social governance.**

# Industrial Capitalism and Surveillance capitalism

The prospect for economic and social improvement offered by AI and Big Data is accompanied by the risks referred to as “**surveillance capitalism**” and the “**surveillance State**”.

According to **Shoshana Zuboff**, **surveillance capitalism** is the **leading economic model** of the present age.

## Industrial Capitalism

**Karl Polanyi** observed that industrial capitalism also treats as **commodities** (products to be sold in the market) **entities that are not produced for the market**: human life becomes “**labour**” to be bought and sold, nature becomes “**land**” or “**real estate**”, exchange becomes “**money**.”

As a consequence, the dynamics of capitalism produces destructive tensions —exploitation, destruction of environment, financial crises— unless **countervailing forces**, such as **law, politics** and **social organisations** (e.g., workers’ and consumers’ movements), intervene to counteract, moderate and mitigate excesses.

# Industrial Capitalism and Surveillance capitalism

The prospect for economic and social improvement offered by AI and Big Data is accompanied by the risks referred to as “**surveillance capitalism**” and the “**surveillance State**”.

According to **Shoshana Zuboff**, **surveillance capitalism** is the **leading economic model** of the present age.

## Surveillance Capitalism

**It expands commodification**, extending it to **human experience**, which it turns into **recorded and analysed behaviour**, i.e., it transforms into **marketable opportunities to anticipate and influence**.

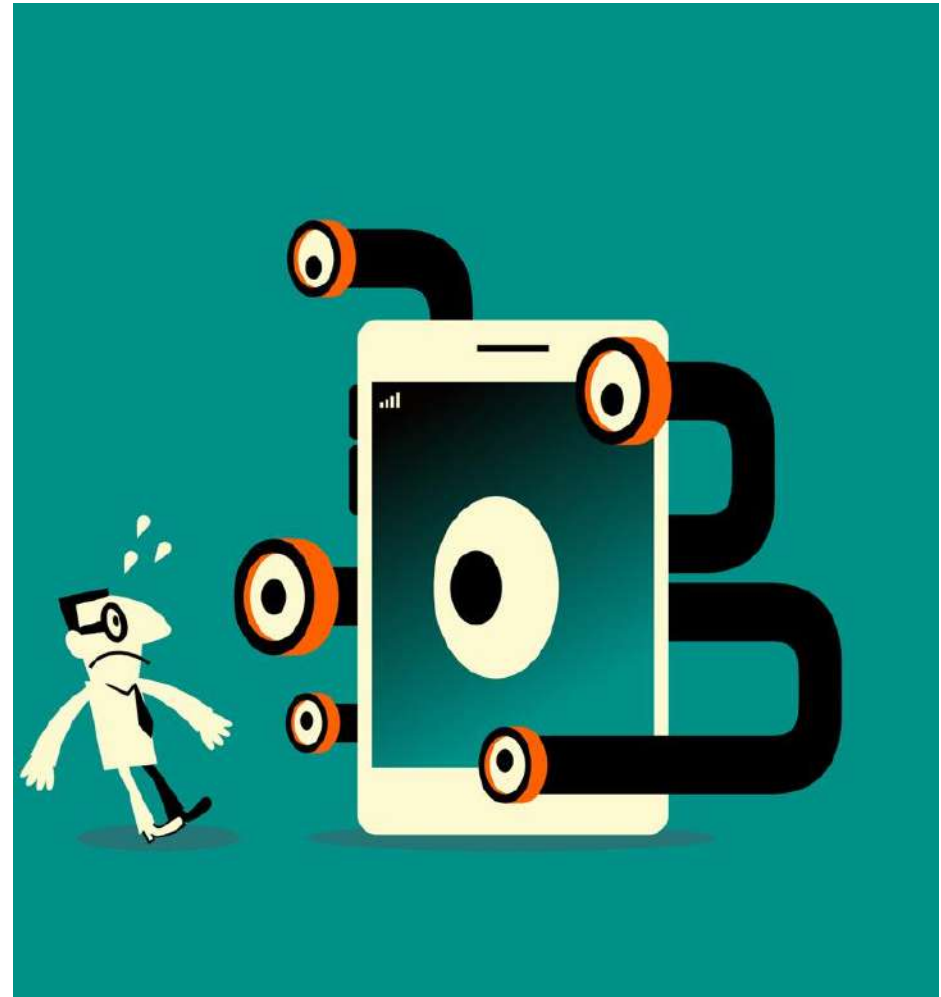
annexes **human experience** to the market dynamic so that it is reborn as behavior: **the fourth “fictional commodity.”**

# Surveillance capitalism

Polanyi's first three **fictional commodities**—land, labor, and money—were subjected to **law**.

Although these laws have been imperfect, the institutions of labor law, environmental law, and banking law are regulatory frameworks intended to defend society (and nature, life, and exchange) from the worst excesses of raw capitalism's destructive power.

**Surveillance capitalism's** expropriation of human experience has faced **no** such **impediments**.

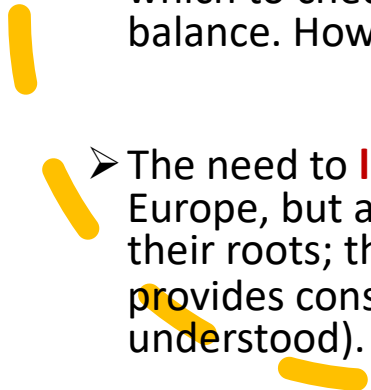






# Surveillance capitalism

- In the case of surveillance capitalism, raw market dynamics can lead to novel disruptive outcomes. Individuals are subject to **manipulation**, are **deprived of control over their future** and **cannot develop their individuality**. Social networks for collaboration are replaced by surveillance-based mechanism of incentives and disincentives. (e.g. Uber recording workers' performance + mutual reviews of workers and clients)
- This new way of governing human behaviour may lead to efficient outcomes, but **it affects the mental wellbeing and autonomy of the individuals concerned**.
- According to Zuboff, we have not yet developed **adequate legal, political or social measures** by which to check the potentially disruptive outcomes of surveillance capitalism and keep them in balance. However, she observes, **the GDPR** could be an important step in this direction
- The need to **limit the commercial use of personal data** has led to new legal schemes not only in Europe, but also in California, the place where many world-leading "surveillance capitalists" have their roots; the CCPA (**California Consumer Privacy Act**), which came into effect on January 2020, provides consumers with rights to access their data and to prohibit data sales (broadly understood).



# Surveillance State

- At the governmental level, surveillance capitalism finds its parallel in the so-called “surveillance State”
- In the National Surveillance State, the government uses surveillance, data collection, collation, and analysis to **identify problems**, to **head off potential threats**, to **govern populations**, and to **deliver valuable social services**.
- The National Surveillance State is a special case of **the Information State**-a state that tries to identify and solve problems of governance through the collection, collation, analysis, and production of information.



# Surveillance State

## Advantage



- Support efficiency in managing public activities
- Coordinate citizens' behaviour
- Prevent social harms

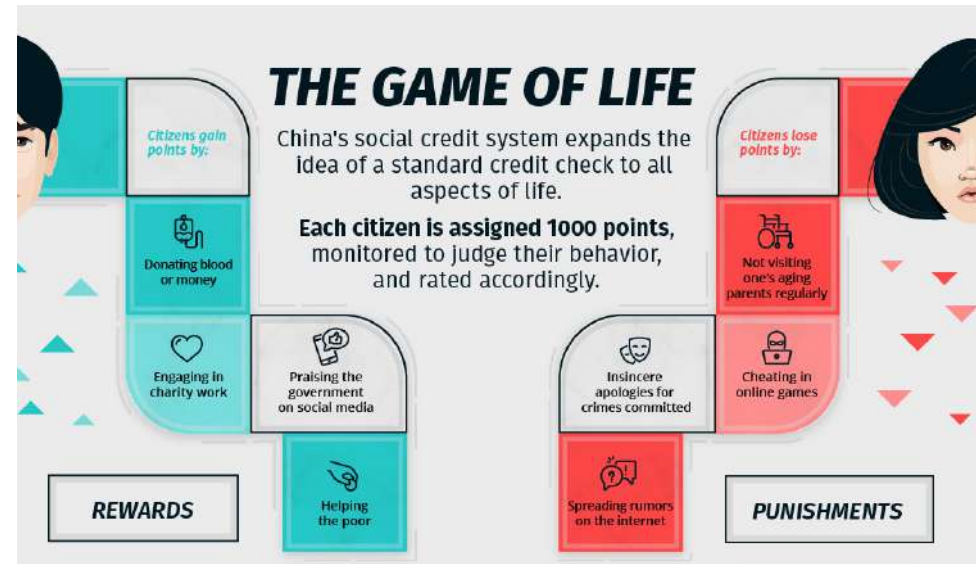
## Disadvantage



- New kinds of influence and control
- Promote values and Purposes that may conflict with democracy
- Diminish autonomy

# Surveillance State: the Chinese Social credit systems

- The Chinese Social credit systems collects data about citizens and assigns to those citizens scores that quantify their social value and reputation.
- It is based on the aggregation and analysis of personal information
- The collected data cover **financial aspects** (e.g., timely compliance with contractual obligations), **political engagement** (e.g., participation in political movements and demonstrations), **involvement in civil and criminal proceedings** (past and present) and **social action** (e.g. participation in social networks, interpersonal relationships, etc.).
- Citizens may be assigned positive or negative points, which contribute to their social score.



- The overall score determines citizens' **access to services and social opportunities**, e.g. universities, housing, transportation, jobs, financing, etc.
- The system's purported **objective** is to **promote mutual trust, and civic virtues**.
- **Risks: opportunism and conformism** may be rather promoted to the detriment of individual autonomy and genuine moral and social motivations.





# Individual and social costs of AI & Big Data applications

In some cases and domain, AI & Big Data applications—even when accurate and unbiased—may have individual and social costs that outweigh their advantages.

- Which systems really deserve to be built?
- Which problems most need to be tackled?
- Who is best placed to build them?
- And who decides?

We need genuine **accountability mechanisms**

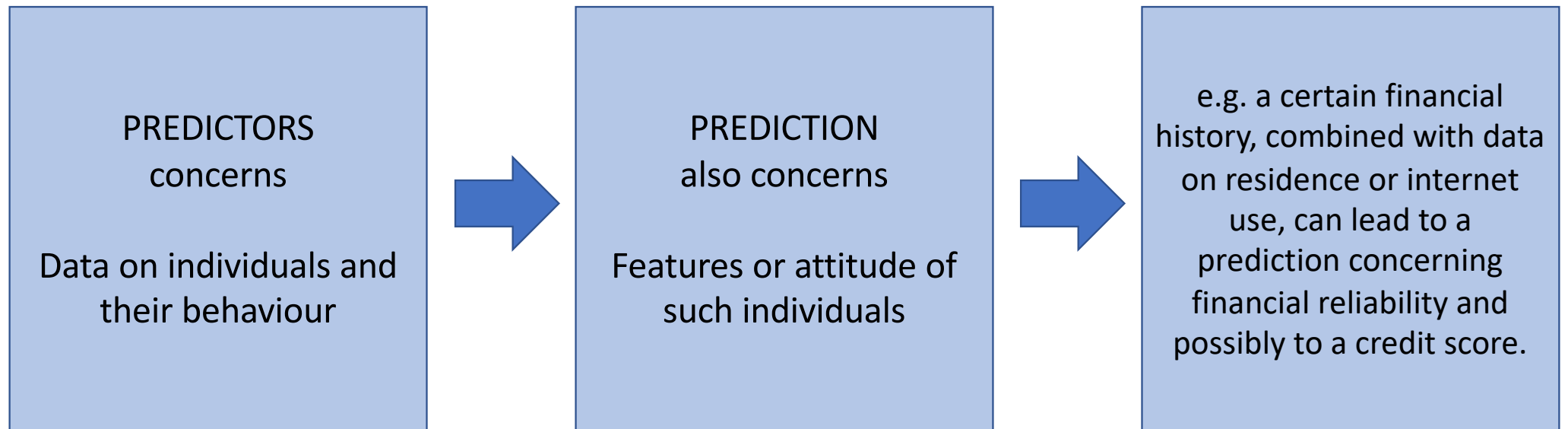
Consider, for instance, systems that are able to recognise sexual orientation, or criminal tendencies from the faces of persons. **Should we just ask whether these systems provide reliable assessments, or should we rather ask whether they should be built at all?**





# The general problem of social sorting and differential treatment

The key aspect of ML systems is the ability to engage in **differential inference: different combinations of predictor-values are correlated to different predictions.**



# The general problem of social sorting and differential treatment

A new dynamic of **stereotyping** and **differentiation** takes place.

(a) The individuals whose data support the same prediction, will be considered and treated in the same way.

(b) The individuals whose data support different predictions, will be considered and treated differently.



This **equalisation and differentiation**, depending on the domains in which it is used and on the purposes that it is meant to serve, may affect **positively or negatively** the individuals concerned.

## Example: use of machine learning technologies to detect or anticipate health issues

- Beneficial application
- **Benefits concern in principle all data subjects** i.e., those) whose data are processed for this purpose
- Processing of health-related data may also be justified on grounds of public health (Article 9 (2)(h)), and in particular for the purpose of “monitoring epidemics and their spread” (Recital 46).
- This provision has become hugely relevant in the context of the Coronavirus disease 2019 (COVID-19) epidemics. In particular a vast debate has been raised by development of **applications for tracing contacts.**



Example: use of machine learning technologies to detect or anticipate health issues

- Such processing should be viewed as **legitimate** as long as it effectively contributes to **limit the diffusion and the harmfulness of the epidemics**, assuming that the **privacy and data protection risks are proportionate to the expected benefit**, and that **appropriate mitigation measures are applied**.

(See the European Data Protection Board Guidelines 04/2020 on the use of location data and contact-tracing tools in the context of the COVID-19 outbreak).



Example: use of the predictions based on health data in the contexts of insurance and recruiting

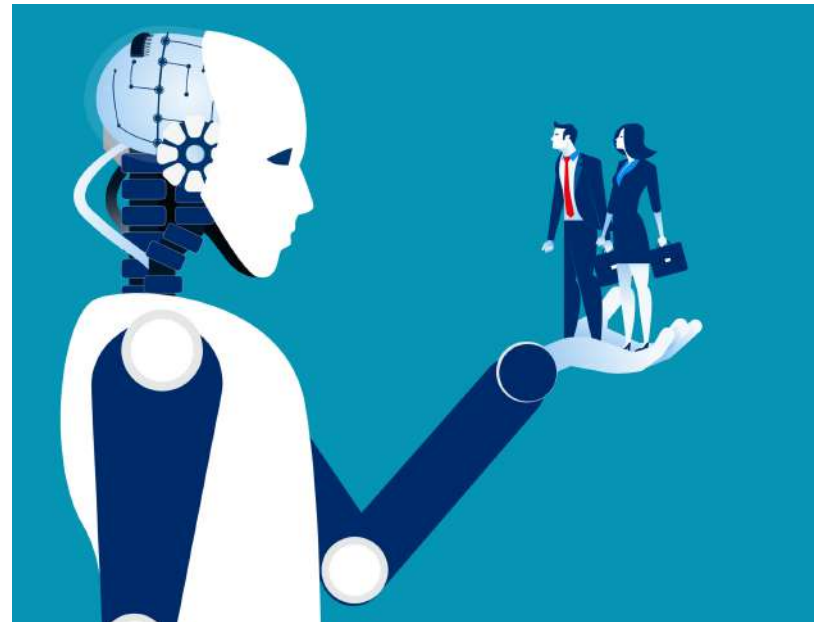
- Predictions based on health data in the context of insurance deserves a **much less favourable assessment**
- **Gainers:** the insured individuals getting a better deal based on their favourable health prospects.
- **Losers:** those getting a worse deal because of their unfavourable prospects.
- Individuals who already are disadvantaged because of their medical conditions would suffer further disadvantage, being excluded from insurance or being subject to less favourable conditions.





Example: use of the predictions based on health data in the contexts of insurance and recruiting

- Insurance companies having the ability to distinguish the risks concerning different applicants would have a **competitive advantage**, being able to provide better conditions to less risky applicants, so that insurers would be pressured to collect as much personal data as possible.
- **Even less commendable would be the use of health predictions in the context of recruiting**, which would involve burdening less healthy people with unemployment or with harsher work conditions. Competition between companies would also be affected, and pressure for collecting health data would grow.



## A further example concerns...

**Price discrimination:** different prices and different conditions to different consumers, depending on predictions on their willingness to pay.

**Risks:**

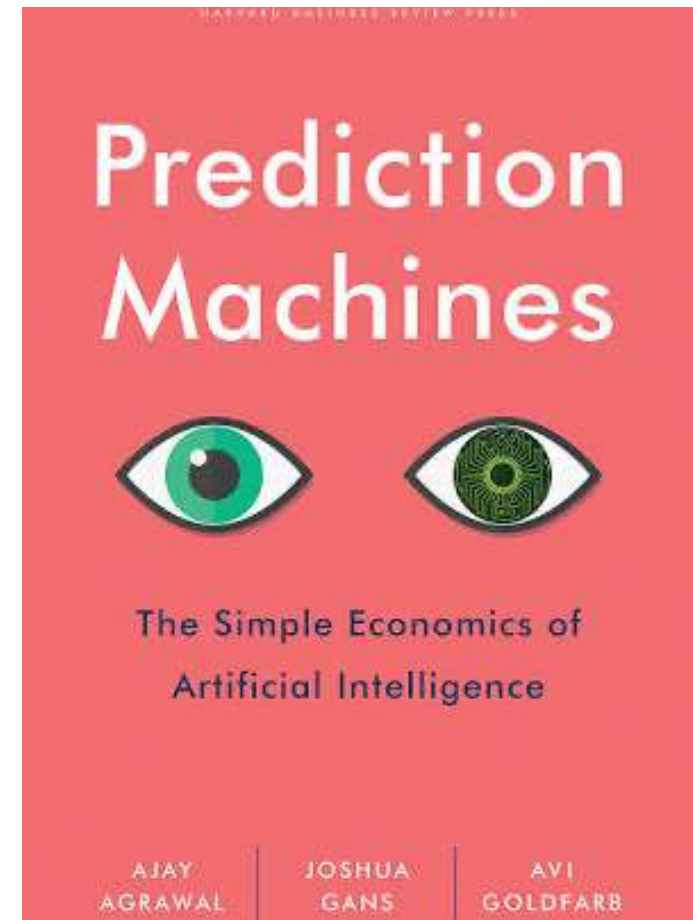
- harm consumers
- individuals may be deprived of access to some opportunities
- affect the functioning of markets
- may be unfair(?)
- undermines the efficiency of the economy



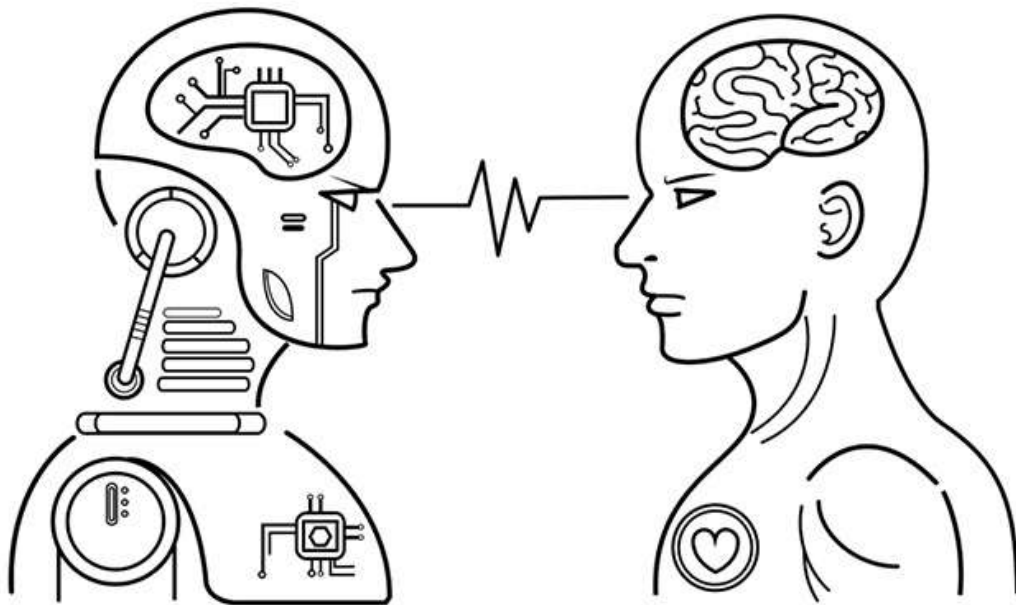
Example: price discrimination

# AI in decision making concerning individuals: fairness and discrimination

- The combination of AI and Big Data enables automated decision-making even in domains that require complex choices, based on multiple factors, and on non-predefined criteria.
- In recent years, a wide debate has taken place on prospects and risks of algorithmic assessments and decisions concerning individuals



# Are AI systems better than humans in assessing us?



In many domains automated predictions and decisions are not only **cheaper**, but also **more precise and impartial** than human ones.

- AI can **avoid typical fallacies of human psychology** (overconfidence, loss aversion, anchoring, confirmation bias, representativeness heuristics, etc.), and the widespread human **inability to process statistical data**, as well as **typical human prejudice** (concerning, e.g., ethnicity, gender, or social background).
- In many assessments and decisions —on investments, recruitment, creditworthiness, or also on judicial matters, such as bail, parole, and recidivism— algorithmic systems have **often performed better**, according to usual standards, than human experts.

# Or not?

Others have underscored the possibility that algorithmic decisions may be **mistaken or discriminatory**.

- Only in rare cases will algorithms engage in explicit unlawful discrimination, so-called **disparate treatment**, basing their outcomes on prohibited features (predictors) such as race, ethnicity or gender.
- More often a system's outcome will be discriminatory due to its **disparate impact**, i.e., since it disproportionately affects certain groups, without an acceptable rationale





# Systems reproducing the strengths and weaknesses of humans in making judgments



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

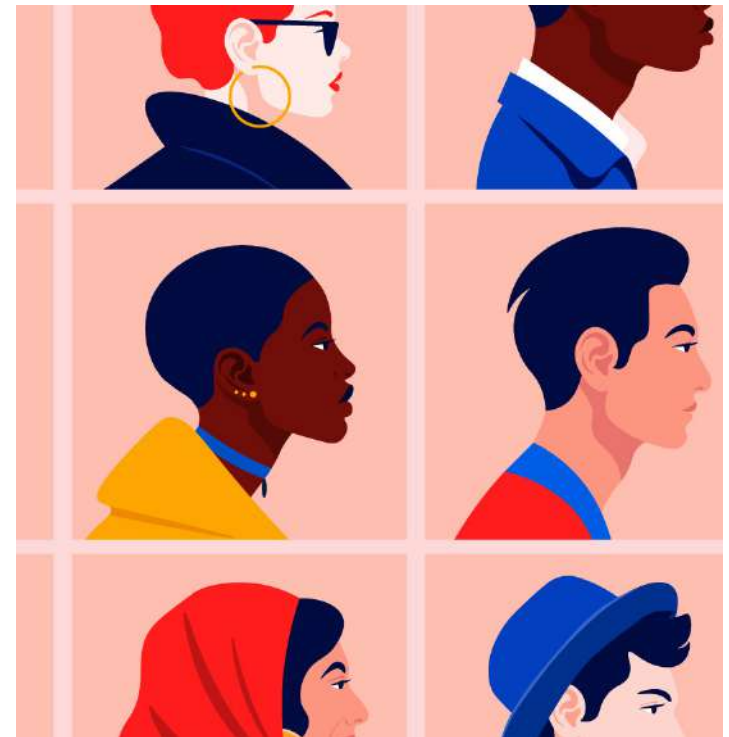
- For example, a recruitment system trained on the past hiring decisions will learn to emulate the managers' assessment of the suitability of candidates, rather than to directly predict an applicant's performance at work. If past decisions were influenced by prejudice, the system will reproduce the same logic.

# Prejudice in the training set

Prejudice baked into training sets may persist even if the inputs (the predictors) to automated systems do not include forbidden discriminatory features (e.g. ethnicity or gender.)

This may happen whenever a **correlation exists between discriminatory features and some predictors**

- Assume, for instance, that a prejudiced human resources manager did not hire applicants from a certain ethnic background, and that people with that background mostly live in certain neighbourhoods. A training set of decisions by that manager will teach the systems not to select people from those neighbourhoods, which would entail continuing to reject applications from the discriminated-against ethnicity. (Kleinberg et al (2019)).



# Systems biased against groups

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

- Assume, for instance, that the **future performance** of employees (the target of interest in job hiring) is only measured by the **number of hours worked in the office**. This outcome criterion will lead to past hiring of women —who usually work for fewer hours than men, having to cope with family burdens— being considered less successful than the hiring of men; based on this correlation (as measured on the basis of the biased proxy), the systems will predict a poorer performance of female applicants.



# System's biases embedded in the predictors

In other cases, mistakes and discriminations may pertain to the machine-learning system's biases embedded in the predictors.

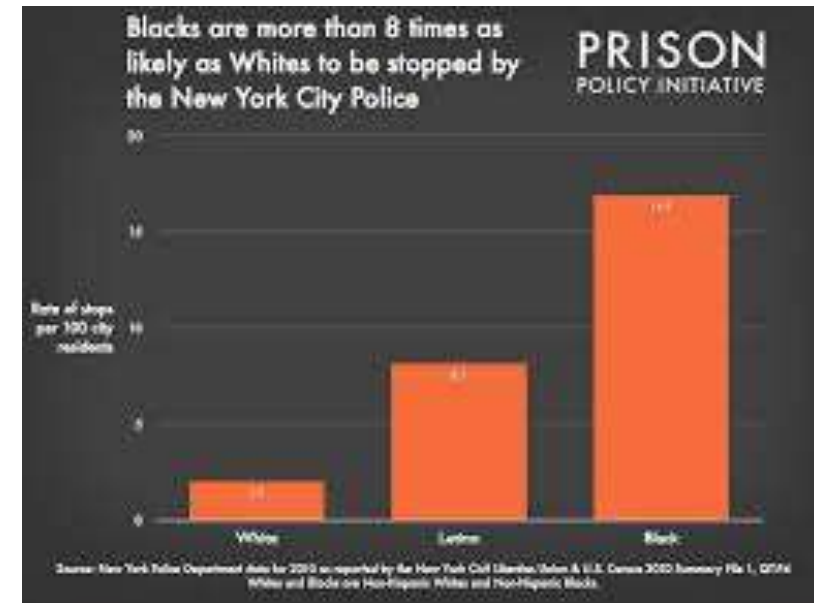
A system may perform unfairly, since it uses a favourable predictor (input feature) that only applies to members of a certain group (e.g., the fact of having attended a socially selective high-education institution).

Unfairness may also result from taking biased human judgements as predictors (e.g., recommendation letters).

# Data set that does NOT reflect the statistical composition of the population

Finally, unfairness may derive from a **data set that does reflect the statistical composition of the population.**

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





- Members of a certain group may also suffer prejudice when that group is only represented by a very small subset of the training set,
- This will reduce the accuracy of predictions for that group (e.g., consider the case of a firm that has appointed few women in the past and which uses its records of past hiring as its training set).



# Challenging the unfairness of automated decision- making

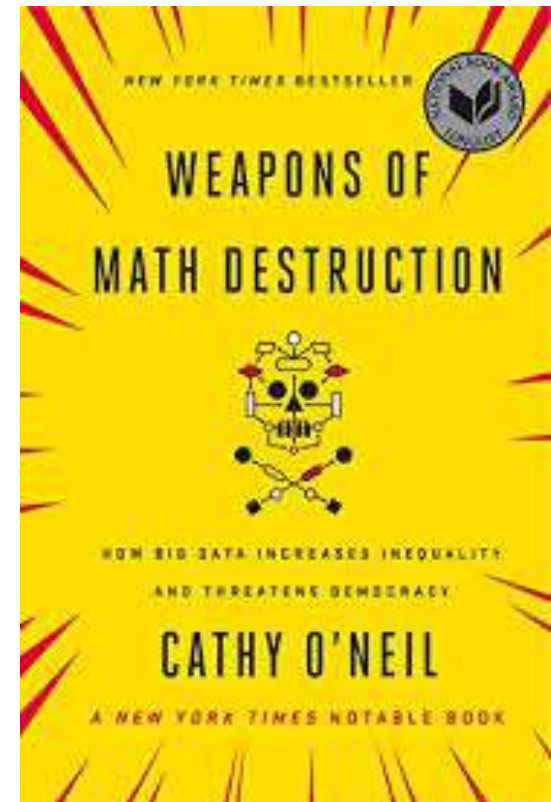
It has been observed that it is difficult to challenge the unfairness of automated decision-making.

Challenges raised by the individuals concerned, even when justified, may be disregarded or rejected because they interfere with the system's operation, giving rise to additional **costs and uncertainties**.

In fact, predictions of machine-learning systems are based on **statistical correlations, against which it may be difficult to argue** on the basis of individual circumstances, even when exceptions would be justified.

# Weapons of math destruction

“An algorithm processes a slew of statistics and comes up with a probability that a certain person might be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone’s life upside down. And yet when the person fights back, “suggestive” countervailing evidence simply won’t cut it. The case must be ironclad. The human victims of WMDs, we’ll see time and again, are held to a far higher standard of evidence than the algorithms themselves”. (O’Neil (2016))



# Or not?

[W]ith appropriate requirements in place, the use of algorithms will make it possible to more easily examine and interrogate the entire decision process, thereby making it far easier to know whether discrimination has occurred. By forcing a new level of specificity, the use of algorithms also highlights, and makes transparent, central **trade-offs among competing values**. Algorithms are not only a threat to be regulated; with the right safeguards in place, they have **the potential to be a positive force for equity**

(Kleinberg, Ludwig, Mullainathan, e Sunstein (2018, 113)).



# Challenging the unfairness of automated decision-making

These criticisms have been countered by observing that **algorithmic systems**, even when based on machine learning, are **more controllable** than human decision-makers, their **faults can be identified** with precision, and **they can be improved** and **engineered to prevent unfair outcomes**.





# Should we exclude the use of automated decision-making?

It seems that issues that have just been presented should not lead us to exclude categorically the use of automated decision-making.

The alternative to automated decision-making is not perfect decisions but human decisions with all their flaws: a biased algorithmic system can still be fairer than an even more biased human decision-maker.

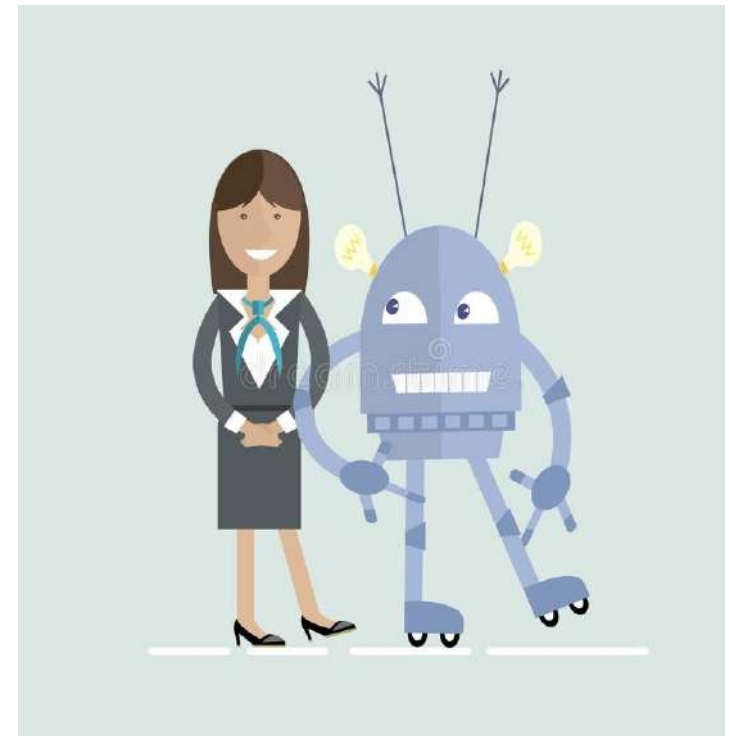


# Humans + Algorithms?

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

In fact, AI systems have demonstrated an ability to successfully also act in domains traditionally entrusted the trained intuition and analysis of humans, e.g., medical diagnosis, financial investment, granting of loans, etc.

The future challenge will consist in finding the best combination between human and AI, taking into account the capacities and the limitations of both.



# Conclusions..

- AI enables new kinds of algorithmic mediated differentiations between individuals
- In the **AI era differential treatments** can be based on **vast amounts of data enabling probabilistic predictions, which may trigger algorithmically predetermined responses.**
- The impacts of such practices can go beyond the individuals concerned, and affect important social institution, in the economical and political sphere.

# Conclusions..

The **GDPR** provides some **constraints**:

- the need for a legal basis for any processing of personal data
- obligations concerning information and transparency
- limitations on profiling and automated decision making
- requirements on anonymisation and pseudonymisation, etc.

These constraints need to be coupled with **strong public oversight**, to ban socially obnoxious forms of differential treatment, and to adopt effective measures that prevent abuses.

## Some interests at stake

- **Interest in data protection and privacy**, namely, the interest in a lawful and proportionate processing of personal data subject to oversight.
- **Interest in fair algorithmic treatment**: concern from an algorithmic transparency/explicability standpoint
- **Individual autonomy**: black box models whose functioning is not accessible and whose decisions remain unexplained and thus unchallengeable.
- **Interest in not being misled or manipulated by AI systems and to trust such systems.**
- **Indirect interest in fair algorithmic competition**, i.e., in not being subject to market-power abuses resulting from exclusive control over masses of data and technologies.



# SOCIAL EMPOWERMENT



## AI technologies for social and legal empowerment

Regulatory instruments and their implementation by public bodies are an essential element but they may be insufficient.

- AI & Big Data are employed in domains already characterized by a **vast power imbalance**, which they may contribute to accentuate.
- The **countervailing power of civil society is needed** to detect abuses, inform the public, activate enforcement, etc.
- In the AI era, an effective **countervailing power needs also to be supported by AI**

# Citizen-empowering technologies - Claudette

Examples of citizen-empowering technologies:

ad-blocking systems

anti-spam software

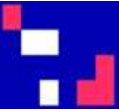
anti-phishing techniques.

A step forward: services deployed with the goal of analysing and summarizing massive amounts of product reviews or comparing prices across a multitude of platforms.

One example in this direction is offered by CLAUDETTE (<https://claudette.eui.eu/>)

# Citizen-empowering technologies – PDA/CDA

- Proposals for automatically extracting, categorizing and summarizing information from **privacy documents**, and assisting users in processing and understanding their contents.
- Multiple AI methods to support data protection could be merged into integrated PDA-CDA (**Privacy digital assistants/consumer digital assistants**), meant to prevent excessive/unwanted/unlawful collection of personal data and well as to protect users from manipulation and fraud, provide them with awareness of fake and untrustworthy information, and facilitate their escape from “**filter bubbles**” (the unwanted filtering/pushing of information).



# AI in the GDPR - outline

- AI in the conceptual framework of the GDPR
- AI and the data protection principles
- AI and legal bases
- AI and transparency
- AI and data subjects' rights
- Automated decision making
- AI and privacy by design



# AI in the conceptual framework of the GDPR

- Unlike the 1995 Data Protection Directive, the **GDPR contains** some terms referring to the **Internet** (Internet, social networks, website, links, etc.), but **it does not contain the term “Artificial Intelligence”**, nor any terms expressing related concepts
- The GDPR is focussed on the **challenges emerging for the Internet** — which were not considered in the 1995 Data Protection Directive, but were well present at the time when GDPR was drafted— rather than on new issues pertaining to AI, which only acquired social significance in most recent years.
- **However, many AI provisions are relevant to GDPR**



## Article 3

# Territorial scope

1. This Regulation applies to the processing of personal data in the context of the activities of an establishment of a controller or a processor in the Union, regardless of whether the processing takes place in the Union or not.

2. This Regulation applies to the processing of personal data of data subjects who are in the Union by a controller or processor not established in the Union, where the processing activities are related to:

(a) the offering of goods or services, irrespective of whether a payment of the data subject is required, to such data subjects in the Union; or

(b) the monitoring of their behaviour as far as their behaviour takes place within the Union.

[...]

# Article 4

## Definitions

- **(1) 'personal data'** means any information relating to an identified or identifiable natural person (**'data subject'**); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
- **(Data subject:** the natural person whom information relates to)
- **(2) 'processing'** means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means,....
- **(7) 'controller'** means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the **purposes and means** of the processing of personal data...
- **(8) 'processor'** means a natural or legal person, public authority, agency or other body which **processes** personal data **on behalf of the controller;**

*Article 5*  
Principles relating to  
processing of  
personal data

- Lawfulness, fairness and transparency
- Purpose limitation
- Data minimisation
- Data accuracy
- Storage limitation
- Integrity and confidentiality
- Accountability principle

# Article 6

## Lawfulness of processing

1. Processing shall be lawful only if and to the extent that at least one of the following applies:

(a) the data subject **has given consent** to the processing of his or her personal data for one or more specific purposes;

(b) processing is **necessary for the performance of a contract** to which the data subject is party or in order to take steps at the request of the data subject prior to entering into a contract;

(c) processing is **necessary for compliance with a legal obligation** to which the controller is subject;

(d) processing is **necessary in order to protect the vital interests** of the data subject or of another natural person;

(e) processing is **necessary for the performance of a task carried out in the public interest** or in the exercise of official authority vested in the controller;

(f) processing is **necessary for the purposes of the legitimate interests pursued by the controller or by a third party**, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of personal data, in particular where the data subject is a child.

2. [...]

# Article 4(1) GDPR: Personal data - identification

Here is how **personal data** are defined in **Article 4 (1) GDPR**:

➤ *‘personal data’ means any information relating to an identified or identifiable natural person (**‘data subject’**); an **identifiable natural person** is one who can be identified, **directly or indirectly**, in particular by reference to an **identifier** such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;*

- What about Natural phenomena?
- What about general medical information on human physiology or pathologies?



# Article 4(1) GDPR: Personal data - identifiability

- **Recital (26)** addresses **identifiability**, namely, the conditions under which a **piece of data which is not explicitly linked to a person, still counts as personal data, since the possibility exists to identify the person concerned.**
- Identifiability depends on the availability of “**means reasonably likely to be used**” for **successful reidentification**, which in its turn, depends on the technological and sociotechnical state of the art:
  - *To determine whether a natural person is identifiable, account should be taken of **all the means reasonably likely to be used**, such as singling out, either by the controller or by another person **to identify the natural person directly or indirectly**. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all **objective factors, such as the costs of and the amount of time required for identification**, taking into consideration the available technology at the time of the processing and technological developments.*

# Article 4(1) GDPR: Personal data - pseudonymisation

- **Pseudonymisation:** the data items that identify a person are substituted with a pseudonym, but **the link between the pseudonym and the identifying data items can be retraced by using separate info** (e.g., a table linking pseudonyms and real names, or through cryptography key to decode the encrypted names)
- Recital (26) specifies that **pseudonymised data still are personal data.**
  - *Personal data which have undergone pseudonymisation, which could be attributed to a natural person by the use of additional information should be considered to be **information on an identifiable natural person.***

# Article 4(1) GDPR: Personal data – connection with technological developments

- The connection between the personal nature of information and **technological development** is mentioned at **Recital (9) of Regulation 2018/1807\***:
  - If technological developments make it possible to turn anonymised data into personal data, such data are to be treated as personal data, and Regulation (EU) 2016/679 is to apply accordingly.
- The concept of **non-personal data** is **not positively defined in the EU legislation**
  - **Examples of non-personal data:** aggregated and anonymised datasets used for Big Data analytics, data on precision farming that can help to monitor and optimise the use of pesticides and water, or data on maintenance needs for industrial machines.”

*\*(Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union)*

# AI and GDPR definition of personal data: Reidentification & further inferences

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

- (1) the “re-personalisation” of anonymous data, namely the **reidentification** of the individuals to which such data are related;
- (2) the **inference** of further personal information from personal data that are already available.

# Reidentification

AI, and methods for **computational statistics**, increases the identifiability of apparently anonymous data, since they enable **nonidentified data** (including data having been anonymised or pseudonymised) to be **connected to the individuals concerned**

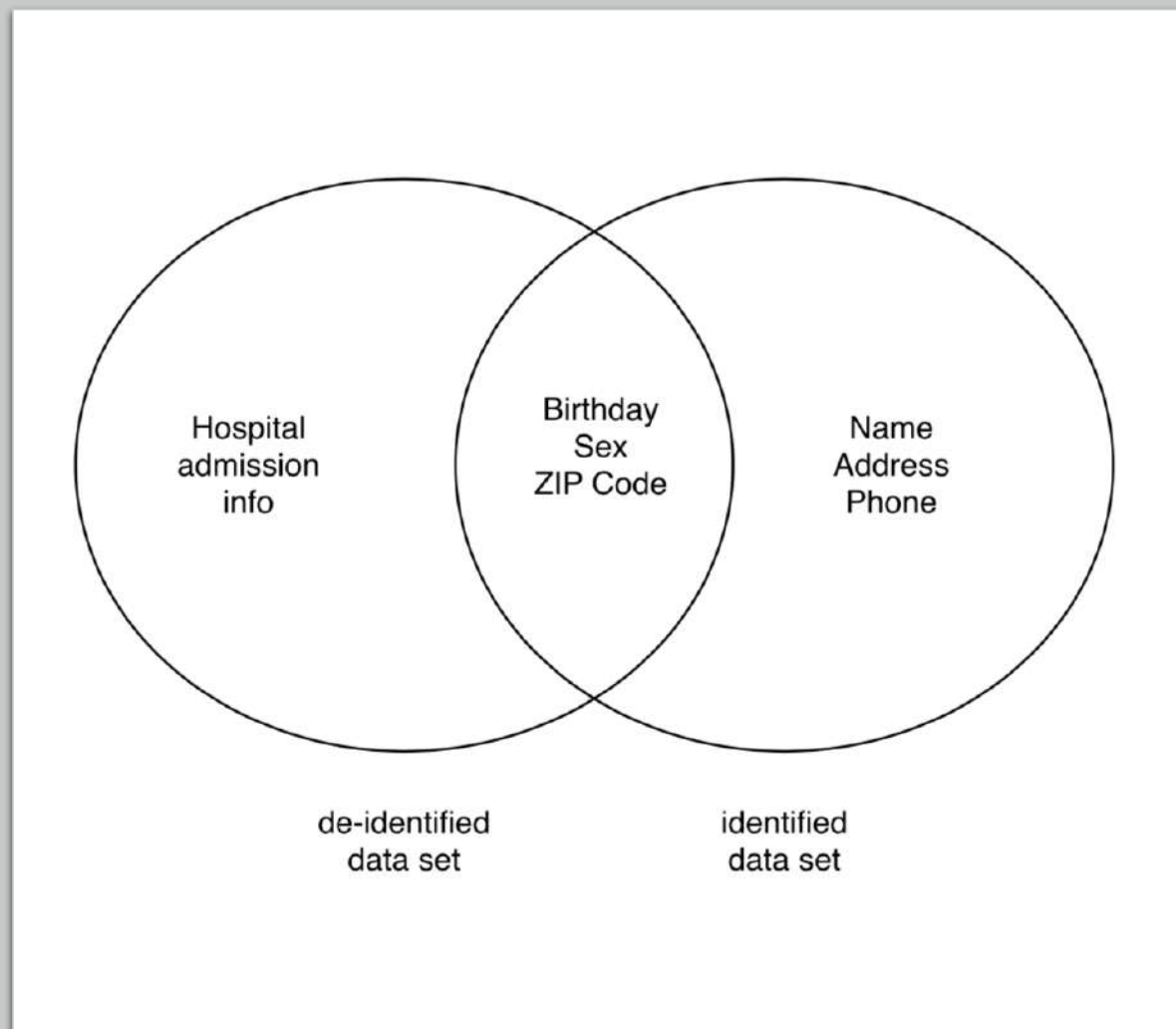
- **[N]umerous supposedly anonymous datasets have recently been released and reidentified.**
  - In 2016, journalists reidentified politicians in an anonymized browsing history dataset of 3 million German citizens, uncovering their medical information and their sexual preferences.
  - A few months before, the Australian Department of Health publicly released de-identified medical records for 10% of the population only for researchers to reidentify them 6 weeks later.
  - Before that, studies had shown that de-identified hospital discharge data could be reidentified using basic demographic attributes and that diagnostic codes, year of birth, gender, and ethnicity could uniquely identify patients in genomic studies data.
  - Finally, researchers were able to uniquely identify individuals in anonymized taxi trajectories in NYC27, bike sharing trips in London, subway data in Riga, and mobile phone and credit card datasets. (Rocher et al 2019).

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



## The connection between identified and de-identified data

The figure illustrates the connection between an identified and a de-identified data set that enabled the reidentification of the health record of the governor of Massachusetts. This result was obtained by searching for de-identified data, such as the information on Hospital admission, that matched the Governor's date of birth, ZIP code and gender.



The connection  
between identified and  
de-identified data

**The Netflix price database case**, in which anonymised movie ratings could be re-identified by linking them to non-anonymous ratings in IMDb (Internet Movie Database). In fact, knowing only two non-anonymous reviews by an IMDb user, it was possible to identify the reviews by the same user in the anonymous database.



# Reidentification

- Reidentification as **a specific kind of inference** of personal data. For an item to be linked to a person, **it is not necessary that the data subject is identified with absolute certainty; a degree of probability may be sufficient**
- Thanks to AI & Big Data the identifiability of the data subjects has vastly increased.
- As it has been argued, **"in any 'reasonable' setting there is a piece of information that is in itself innocent, yet in conjunction with even a modified (noisy) version of the data yields a privacy breach."**

This possibility can be addressed in two ways:

1. The first consists in ensuring that data is deidentified in ways that **make it more difficult to reidentify** the data subject;
2. The second consists in **implementing security processes and measures** for the release of data that contribute to this outcome.

# Inferred personal data

- AI systems may **infer new information** about data subjects, by applying algorithmic models to their personal data.
- The key issue is **whether the inferred information should be considered as new personal data**, distinct from the data from which it has been inferred.
  - Assume for instance, that an individual's sexual orientation is inferred from his or her facial features or that an individual's personality type is inferred from his or her online activity. Is the inferred sexual orientation or personality type a new item of personal data? Even when the inference only is probabilistic?
- If the inferred information counts as new personal data, then automated inferences would trigger all the consequences that the processing of personal data entails according to the GDPR.

# Legal status of automatically inferred information

- Some clues on the legal status of automatically inferred information can be obtained by considering the status of information inferred by humans: there is **uncertainty about whether assertions concerning individuals, resulting from human inferences and reasoning may be regarded as personal data.**
- This issue has been examined by the **ECJ** in **Joint Cases C-141 and 372/12**, where it was denied that the legal analysis, by the competent officer, on an **application for a residence permit** could be deemed personal data. **According to the ECJ, only the data on which the analysis was based** (the input data about the applicant) **as well as the final conclusion** of the analysis (the holding that the application was to be denied) **were to be regarded as personal data.**
- This qualification did not apply to the intermediate steps (the intermediate conclusions in the argument chain) leading to the final conclusion.



# Legal status of automatically inferred information

- In the subsequent decision on **Case C-434/16**, concerning a candidate's request to exercise data protection rights relative to an **exam script and the examiners' comments**, the ECJ apparently departed from the principle stated in **Joint Cases C-141 and 372/12**, arguing that **the examiner's comments, too, were personal data**.
- However, **the Court held that data protection rights, and in particular the right to rectification, should be understood in connection with the purpose of the data at issue**. Thus, according to the Court, **the right to rectification does not include a right to correct a candidate's answers or the examiner's comments (unless they were incorrectly recorded)**.
- In fact, according to the ECJ, data protection law is not intended to ensure the accuracy of decision-making processes or good administrative practices. Thus, an examinee has the right to access both to the exam data (the exam responses) and the reasoning based on such data (the comments), but **he or she does not have a right to correct the examiners' inferences (the reasoning)** or the final result.

The view that **inferred data are personal data** was endorsed by the **Article 29 WP** (Opinion 4/2007)

- *in case of automated inference (profiling) data subjects have the right to access both the input data and the (final or intermediate) conclusions automatically inferred from such data.*

## Article 4(2) GDPR: Profiling

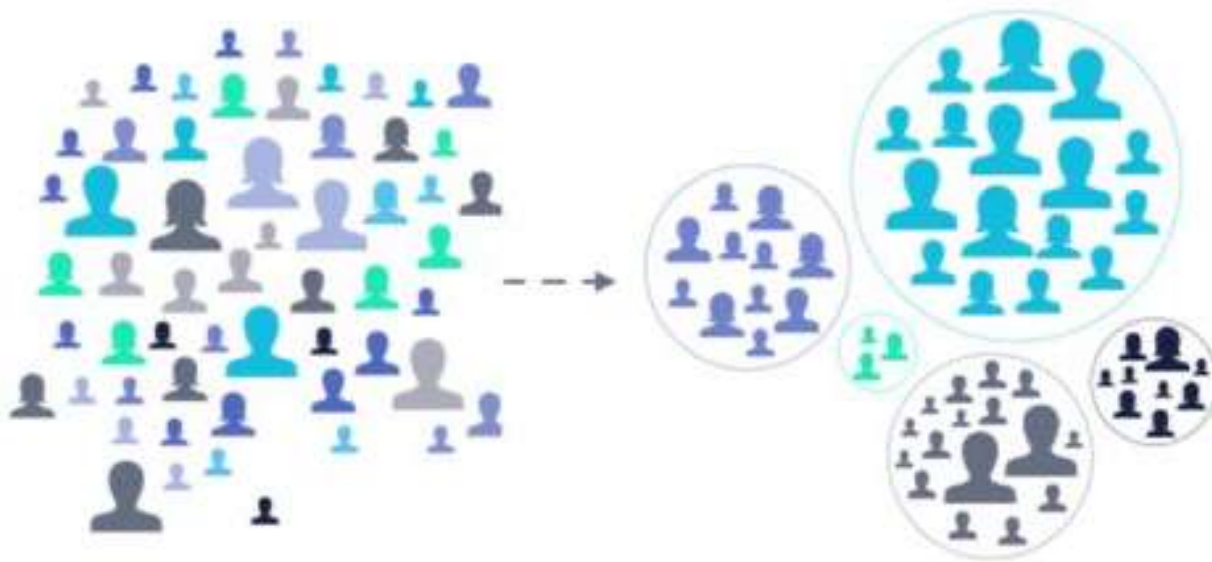
The definition of profiling, while not explicitly referring to AI, addresses processing that today is typically accomplished using AI technologies. This processing consists in using the data concerning person to infer information on further aspects of that person:

*'profiling' means any form of **automated processing of personal data** consisting of the use of personal data to evaluate certain **personal aspects** relating to a natural person, in particular to **analyse or predict aspects** concerning that **natural person's** performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location or movements*



# Article 4(2) GDPR: Profiling

## Segmentation and Profiling



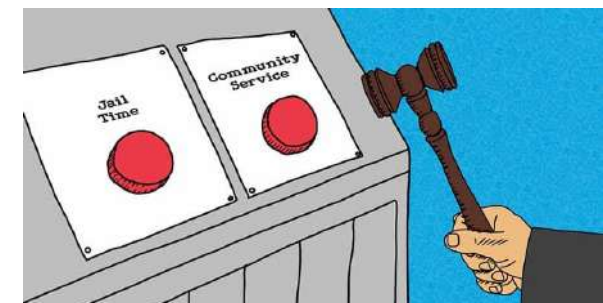
According to the **Article 29 WP**, profiling aims at **classifying persons into categories of groups sharing the features being inferred** (Opinion 216/679):

“broadly speaking, profiling means ***gathering information about an individual*** (or group of individuals) and evaluating their characteristics or behaviour patterns in order to place them into a certain category or group, in particular to analyse and/or make predictions about, for example, their ability to perform a task, interests or likely behaviour.”

# AI and profiling

- **AI & Big Data have vastly increased the opportunities for profiling.**
- Assume that a classifier has been trained on a vast set of past examples, which **link certain features** of individuals (the **predictors**), to **another feature** of the same individuals (the **target**).
- Through the training, the system has learned an algorithmic model that can be applied to new cases: **if the model is given predictors-values concerning a new individual, it infers a corresponding target value for that individual, i.e., a new data item concerning him or her.**
  - the likelihood of heart disease of applicants for insurance on the basis of their health records, their habits or social conditions;
  - the creditworthiness of loan applicants on the basis of their financial history, their online activity and social condition;
  - the likelihood that convicted persons may reoffend on the basis their criminal history, their character (as identified by personality test) and personal background.

These predictions may trigger automated determinations concerning, respectively, the price of a health insurance, the granting of a loan, or the release on parole.



# AI and profiling

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

- Examples: trigger the desired purchasing behaviour, or the desired voting behaviour.





# Inferences as personal data

We need to **distinguish the general correlations that are captured by the learned algorithmic model, and the results of applying that model** to the description of a particular individual.

- *Consider for instance a machine learning system that has learned a model (e.g., a neural network or a decision tree) from a training set consisting of previous **loan applications and outcomes**. The system's training set consists of personal data: e.g., for each borrower, his name, the data collected on him or her —age, economic condition, education, job, etc.— and the information on whether he or she defaulted on the loan.*
- The learned algorithmic model no longer contains personal data, since it links any possible combinations of possible input values (predictors) to a corresponding likelihood of default (target). **The correlations embedded in the algorithmic model are not personal data, since they apply to all individuals sharing similar characteristics. We can possibly view them as group data, concerning the set of such individuals** (e.g., those who are assigned a higher likelihood of default, since they have a low revenue, live in a poor neighbourhood, etc.).
- **Assume that the algorithmic model is then applied** to the input data consisting in the description of a new applicant, in order to determine that applicant's risk of default. **In this case both the description of the applicant and the default risk attributed to him or her by the model represent personal data**, the first being collected data, and the second inferred data.

# Rights over inferences: access

Since **inferred data concerning individuals also are personal data** under the GDPR —at least when they are used to derive conclusions that are or may be acted upon— **data protection rights should in principle also apply**, though concurrent remedies and interests have to be taken into account.

According to the Article 29 Working Party, in the case of automated inferences (profiling) **data subjects have a right to access** both the personal data used as **input** for the inference, and the personal data obtained as (final or intermediate) **inferred output**.



# Rights over inferences: rectification

On the contrary, **the right to rectification only applies to a limited extent**. When the data are processed by a public authority, it should be considered whether review procedures already exist which provide for access and control. In the case of processing by private controllers, the right to rectify the data should be balanced with the respect for autonomy of private assessments and decisions.



According to the Article 29 Working Party data subjects **have a right to rectification** of inferred information not only when the inferred information is “verifiable” (its correctness can be objectively determined), but also when it is the outcome of unverifiable or probabilistic inferences (e.g., a the likelihood of developing heart disease in the future).

In the latter case, rectification may be needed not only when the statistical inference was mistaken, but also when the data subject provides specific additional data that support a different, more specific, statistical conclusion. This is linked to the fact that statistical inferences concerning a class may not apply to subclasses of it

# A general right to “reasonable inference”?

Legal scholars have argued that data subjects should be granted a general right to “reasonable inference” i.e., the right that any assessment of decision affecting them is obtained through automated inferences that are reasonable, respecting both **ethical and epistemic standards**.

Data subject should be entitled to challenge the inferences (e.g. credit scores) made by an AI system, and not only the decisions based on such inferences (e.g., the granting of loans). **It has been argued that for an inference to be reasonable it should satisfy the following criteria:**

- a) **Acceptability:** the input data (the predictors) for the inference should be normatively acceptable as a basis for inferences concerning individuals (e.g., to the exclusion of prohibited features, such as sexual orientation);
- b) **Relevance:** the inferred information (the target) should be relevant to the purpose of the decision and normatively acceptable in that connection (e.g., ethnicity should not be inferred for the purpose of giving a loan).
- c) **Reliability:** both input data, including the training set, and the methods to process them should be accurate and statistically reliable

# A general right to “reasonable inference”?

**Controllers, conversely, should be prohibited to base their assessment or decisions on unreasonable inferences,** and they should also have the obligation to demonstrate the reasonableness of their inferences.

The idea that unreasonable automated inference should be prohibited only applies to inferences meant to lead to **assessments and decisions affecting the data subject**. They should not apply to inquiries that are motivated by merely cognitive purposes, such as those pertaining to scientific research.



# Consent

## **Art 4(11)**

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

## **Art 7 (Conditions for consent)**

1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.
2. If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.
3. The data subject shall have the right to withdraw his or her consent at any time. The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. Prior to giving consent, the data subject shall be informed thereof. It shall be as easy to withdraw as to give consent.
4. When assessing whether consent is freely given, utmost account shall be taken of whether, inter alia, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.

## Information to be provided to the data subject (art 13-14, recital 42 GDPR, art29WP Guidelines on consent)

- **Identity of the controller** and (where applicable) controller's representative, + their contact details
- **Contact details of the data protection officer**
- **Purposes** of the processing for which the personal data are intended
- **Legal basis** for the processing
- **Categories of personal data concerned**
- **Recipients or categories of recipients** of the personal data
- **Period** for which the personal data will be stored, or if that is not possible, the criteria used to determine that period
- Existence of the right to request from the controller **access to and rectification** or **erasure** of personal data or restriction of processing concerning the data subject and to object to processing as well as **the right to data portability**
- **Right to lodge a complaint** with a supervisory authority
- **Source** from which personal data originate
- **Existence of automated decision-making**, including **profiling**

# Article 17

## Right to erasure ('right to be forgotten') (1/2)

1. The data subject shall have the **right to obtain** from the controller the **erasure** of personal data concerning him or her **without undue delay** and the controller shall have the obligation to erase personal data without undue delay where one of the following grounds applies:

- (a) the personal data are no longer necessary in relation to the purposes for which they were collected or otherwise processed;
- (b) the data subject **withdraws consent** on which the processing is based according to point (a) of Article 6(1), or point (a) of Article 9(2), and where there is **no other legal ground for the processing**;
- (c) the data subject **objects to the processing** pursuant to Article 21(1) and there are **no overriding legitimate grounds** for the processing, or the data subject objects to the processing pursuant to Article 21(2);
- (d) the personal data have been **unlawfully processed**;
- (e) the personal data have to be erased for **compliance with a legal obligation** in Union or Member State law to which the controller is subject;
- (f) the personal data have been collected in relation to the offer of information society services referred to in Article 8(1).
- [...]



# Article 17

## Right to erasure ('right to be forgotten') (2/2)

2. Where the controller has made the personal data public and is obliged pursuant to paragraph 1 to erase the personal data, the controller, taking account of available technology and the cost of implementation, shall take reasonable steps, including technical measures, to inform controllers which are processing the personal data that the data subject has requested the erasure by such controllers of any links to, or copy or replication of, those personal data.

3. **Paragraphs 1 and 2 shall not apply to the extent that processing is necessary:**

**(a) for exercising the right of freedom of expression and information;**

**(b) for compliance with a legal obligation** which requires processing by Union or Member State law to which the controller is subject or for the performance of a task carried out in the **public interest** or in the exercise of official authority vested in the controller;

**(c) for reasons of public interest in the area of public health** in accordance with points (h) and (i) of Article 9(2) as well as Article 9(3);

**(d) for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes** in accordance with Article 89(1) in so far as the right referred to in paragraph 1 is likely to render impossible or seriously impair the achievement of the objectives of that processing; or

**(e) for the establishment, exercise or defence of legal claims.**

## Article 9

### Processing of special categories of personal data (1/2)

- 1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be **prohibited**.





# Article 9

## Processing of special categories of personal data (2/2)

2. Paragraph 1 shall not apply if one of the following applies:

(a) the data subject has given **explicit consent** to the processing of those personal data for one or more specified purposes...

(b) processing is necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of **employment** and **social security** and **social protection** law...

(c) processing is necessary **to protect the vital interests of the data subject** or of another natural person where the data subject is physically or legally incapable of giving consent...

(d) processing is carried out in the course of its **legitimate activities** with appropriate safeguards by a foundation, association or any other not-for-profit body with a political, philosophical, religious or trade union aim and on condition that the processing relates solely to the members...

(e) processing relates to personal data which are **manifestly made public by the data subject**;

(f) processing is necessary for **the establishment, exercise or defence of legal claims** or whenever courts are acting in their judicial capacity;

(g) processing is necessary for reasons of **substantial public interest**...

(h) processing is necessary for the purposes of **preventive or occupational medicine**...

(i) processing is necessary for reasons of **public interest in the area of public health**...

(j) processing is necessary for **archiving purposes in the public interest, scientific or historical research purposes or statistical purposes**...

## Article 22

### Automated individual decision-making, including profiling

1. 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 or similarly significantly affects** him or her.
  
2. Paragraph 1 shall not apply if the decision:
  - (a) is necessary for entering into, or performance of, a contract between the data subject and a data controller;
  - (b) is authorised by Union or Member State law to which the controller is subject and which also lays down suitable measures to safeguard the data subject's rights and freedoms and legitimate interests; or
  - (c) is based on the data subject's explicit consent.

# Article 22(1) GDPR: The prohibition of automated decisions

- The first paragraph of Article 22 provides for a general right not to be subject to completely automated decisions significantly affecting the data subject:

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 or similarly significantly affects him or her.
- According to the Article 29 Working Party:

as a rule, there is a **general prohibition** on fully automated individual decision-making, including profiling that has a legal or similarly significant effect.
- For the application of the prohibition established by Article 22(1), four conditions are needed:
  - (1) a decision must be taken
  - (2) it must be solely based on automated processing
  - (3) it must include profiling
  - (4) it must have legal or anyway significant effect.

# Article 22(1) GDPR: conditions for the prohibition of automated decisions

- (1) **a decision must be taken:** requires that a stance be taken toward a person, and that this stance is likely to be acted upon (as when assigning a credit score).
- (2) **it must be solely based on automated processing:** requires that humans do not exercise any real influence on the outcome of a decision-making process, even though the final decision is formally ascribed to a person. This condition is not satisfied when the system is only used as a decision-support tool for humans
- (3) **it must include profiling:** requires that the automated processing determining the decision includes profiling. (A different interpretation of the condition may be suggested, but Recital (71) seems to confirm the first interpretation)
- (4) **it must have legal or anyway significant effect:** Recital (71) mentions the following examples of decision having significant effects: the “automatic refusal of an online credit application or e-recruiting practices”. It has been argued that such effects cannot be merely emotional, and that usually they are not caused by targeted advertising, unless “advertising involves blatantly unfair discrimination in the form of web-lining and the discrimination has non-trivial economic consequences

# Article 21 (1) and (2): Objecting to profiling and direct marketing

- Article 21 (1) specifies that the right to object also applies to profiling:
  - The data subject shall have the right to object, on grounds relating to his or her particular situation, at any time to processing of personal data concerning him or her which is based on point (e) or (f) of Article 6(1), including profiling based on those provisions.
- Profiling in the context of direct marketing is addressed in Article 21 (2), which recognises an unconditioned right to object:
  - Where personal data are processed for direct marketing purposes, the data subject shall have the right to object at any time to processing of personal data concerning him or her for such marketing, which includes profiling to the extent that it is related to such direct marketing.
- This means that the data subject does not need to invoke specific grounds when objecting to processing for direct marketing purposes, and that such purposes cannot be “compelling legitimate grounds for the processing which override the interests, rights and freedoms of the data subject”.
- Given the importance of profiling for marketing purposes, the unconditional right to object to such processing is particularly significant for the self-protection of data subjects. Controllers should be required to provide easy, intuitive and standardised ways to facilitate the exercise of this right.



# Information on automated decision making

Article 13(2)(f) and 14(2)(g) GDPR address a key aspect of AI applications, i.e. automated decision making. The controller has the obligation to provide:

- (a) information on “**the existence of automated decision-making**, including profiling, referred to in Article 22(1)” and
- (b) “at least in those cases meaningful information about **the logic involved**, as well as the significance and the envisaged **consequences** of such processing for the data subject.”

# Information on automated decision making

- **Computer scientists** have focused on the technological possibility of providing understandable models of opaque AI systems (and, in particular, of deep neural networks), i.e., model of the functioning of such systems that can be mastered by human experts. For instance, the following kinds of explanations are at the core of current research on explainable AI:
  - **Model explanation**, i.e., the global explanation of an opaque AI system through an interpretable and transparent model that fully captures the logic of the opaque system.  
This would be obtained for instance, if a decision tree or a set of rules was provided, whose activation exactly (or almost exactly) reproduces the functioning of a neural network.
  - **Model inspection**, i.e., a representation that makes it possible to understanding of some specific properties of an opaque model or of its predictions.  
It may concern the patterns of activation in the system's neural networks, or the system's sensitivity to changes in its input factors (e.g. how a change in the applicant's revenue or age makes a difference in the grant of a loan application).
  - **Outcome explanation**, i.e., an account of the outcome of an opaque AI in a particular instance.  
For instance, a special decision concerning an individual can be explained by listing the choices that lead to that conclusions in a decision tree (e.g., the loan was denied because of the applicant's income fell below a certain threshold)
- The explanatory techniques and models developed within computer science are intended for technological experts and assume ample access to the system being explained.

# Information on automated decision making

**Social scientists** have focused on the objective of making explanations accessible to lay people, thus addressing the communicative and dialectical dimensions of explanations. For instance, it has been argued that the following approaches are needed (Miller 2019, Mittelstadt and Wachter 2019).

- **Contrastive explanation:** specifying what input values made a difference, determining the adoption of a certain decision (e.g., refusing a loan) rather than possible alternatives (granting the loan);
- **Selective explanation:** focusing on those factors that are most relevant according to human judgement;
- **Causal explanation:** focusing on causes, rather than on merely statistical correlations (e.g., a refusal of a loan can be causally explained by the financial situation of the applicant, not by the kind of Facebook activity that is common for unreliable borrowers);
- **Social explanation:** adopting an interactive and conversational approach in which information is tailored according to the recipient's beliefs and comprehension capacities.

While these suggestions are useful for the ex-post explanation of specific decisions by a system, they cannot be easily applied ex-ante, at the time of data collection (or repurposing).

# Information on automated decision making

- Ex-ante the user should ideally be provided with the following information:
- The **input data** that the system takes into consideration (e.g., for a loan application, the applicant's income, gender, assets, job, etc.), and **whether different data items are favouring or rather disavouring the outcome** that the applicant hopes for;
- **The target values** that the system is meant to compute (e.g., a level of creditworthiness, and possibly the threshold to be reached in order for the loan to be approved);
- **The envisaged consequence of the automated assessment/decision** (e.g., the approval or denial of the loan application).
- It may also be useful to specify what are **the overall purposes** that the system is aimed to achieve

# A right to explanation?

According to Recital (71), the safeguards to be provided to data subjects in case of automated decisions include all of the following:

- specific information
- the right to obtain human intervention,
- the right to express his or her point of view,
- the right to obtain an explanation of the decision reached after such assessment
- the right to challenge the decision.

According to Article 22 the suitable safeguards to be provided include “at least”

- the right to obtain human intervention,
- the right to express his or her point of view,
- the right to challenge the decision.

Thus, two items are missing in article 22 relative to Recital (71): the provision of “specific information” and the right to *obtain an explanation of the decision reached after such assessment*”.

**The second omission in particular raises the issue of whether controllers are really required by law to provide an individualised explanation**



# A right to explanation? Two possible interpretations

- **According to the first interpretation**, the European legislator, by only including the request for specific explanation in the recitals and omitting it from the articles of the GDPR, intended to convey a double message: **to exclude an enforceable legal obligation to provide individual explanations**, while recommending that data controllers provide such explanations when convenient, according to their discretionary determinations.
- **Following this interpretation, providing individualised explanation would only be a good practice, and not a legally enforceable requirement.**

# A right to explanation? Two possible interpretations

- **According to the second interpretation**, the European legislator intended on the contrary to **establish an enforceable legal obligation to provide individual explanation**, though without unduly burdening controllers.
- This interpretation is hinted at by the qualifier “at least”, which precedes the reference made to a “right to obtain human intervention on the part of the controller, to express his or her point of view and to contest the decision.” The qualifier seems to suggest that some providers are legally required to adopt further safeguards, possibly including individualised explanations, as indicated in Recital 71.
- **On this second approach, an explanation would be legally needed**, whenever it is practically possible, i.e., whenever it is compatible with technologies, costs, and business practices.
- However, **we should be cautioned against overemphasising a right to individualised explanations as a general remedy to the biases, malfunctions, and inappropriate applications of AI & Big Data technologies** : the right to an explanation is likely to remain underused by the data subjects, given that they may lack a sufficient understanding of technologies and applicable normative standards. Moreover, even when an explanation elicits potential defects, the data subjects may be unable to obtain a new, more satisfactory decision.

# Article 25

## Data protection by design and by default

1. Taking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing, **the controller shall**, both at the time of the determination of the means for processing and at the time of the processing itself, **implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation**, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects.

2. **The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed.** That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.

[...]

# Article 32

## Security of processing

1. Taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk of varying likelihood and severity for the rights and freedoms of natural persons, the controller and the processor shall implement appropriate technical and organisational measures to ensure a level of security appropriate to the risk, including inter alia as appropriate:

(a) the pseudonymisation and encryption of personal data;

(b) the ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services;

(c) the ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident;

(d) a process for regularly testing, assessing and evaluating the effectiveness of technical and organisational measures for ensuring the security of the processing.

[...]

# European Data Protection Board and European Data Protection Supervisor

## Article 68

### European Data Protection Board

1. The European Data Protection Board (the 'Board') is hereby established as a body of the Union and shall have legal personality.
  2. The Board shall be represented by its Chair.
  3. The Board shall be composed of the head of one supervisory authority of each Member State and of the European Data Protection Supervisor, or their respective representatives.
  4. Where in a Member State more than one supervisory authority is responsible for monitoring the application of the provisions pursuant to this Regulation, a joint representative shall be appointed in accordance with that Member State's law.
  5. The Commission shall have the right to participate in the activities and meetings of the Board without voting right. The Commission shall designate a representative. The Chair of the Board shall communicate to the Commission the activities of the Board.
- [...]

## Article 70

### Tasks of the Board

1. The Board shall ensure the consistent application of this Regulation. To that end, the Board shall, on its own initiative or, where relevant, at the request of the Commission, in particular:
  - (a) monitor and ensure the correct application of this Regulation in the cases provided for in Articles 64 and 65 without prejudice to the tasks of national supervisory authorities;
  - (b) advise the Commission on any issue related to the protection of personal data in the Union, including on any proposed amendment of this Regulation;

[...]

  - (e) examine, on its own initiative, on request of one of its members or on request of the Commission, any question covering the application of this Regulation and issue guidelines, recommendations and best practices in order to encourage consistent application of this Regulation;

[...]

# AI Ethics at IBM: From Principles to Practice

Francesca Rossi

IBM fellow and AI Ethics Global Leader

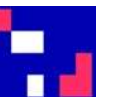
**MAI4CAREU**

Master programmes in Artificial  
Intelligence 4 Careers in Europe



Co-financed by the European Union  
Connecting Europe Facility

This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423





# A brief history of AI

## ARTIFICIAL INTELLIGENCE

Intelligent algorithms defined and coded by people into machines



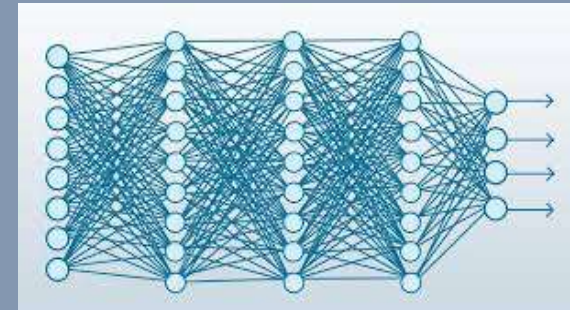
## MACHINE LEARNING

Ability to learn without being explicitly programmed



## DEEP LEARNING

Learning based on Deep Neural Networks



1950's

1960's

1970's

1980's

1990's

2000's

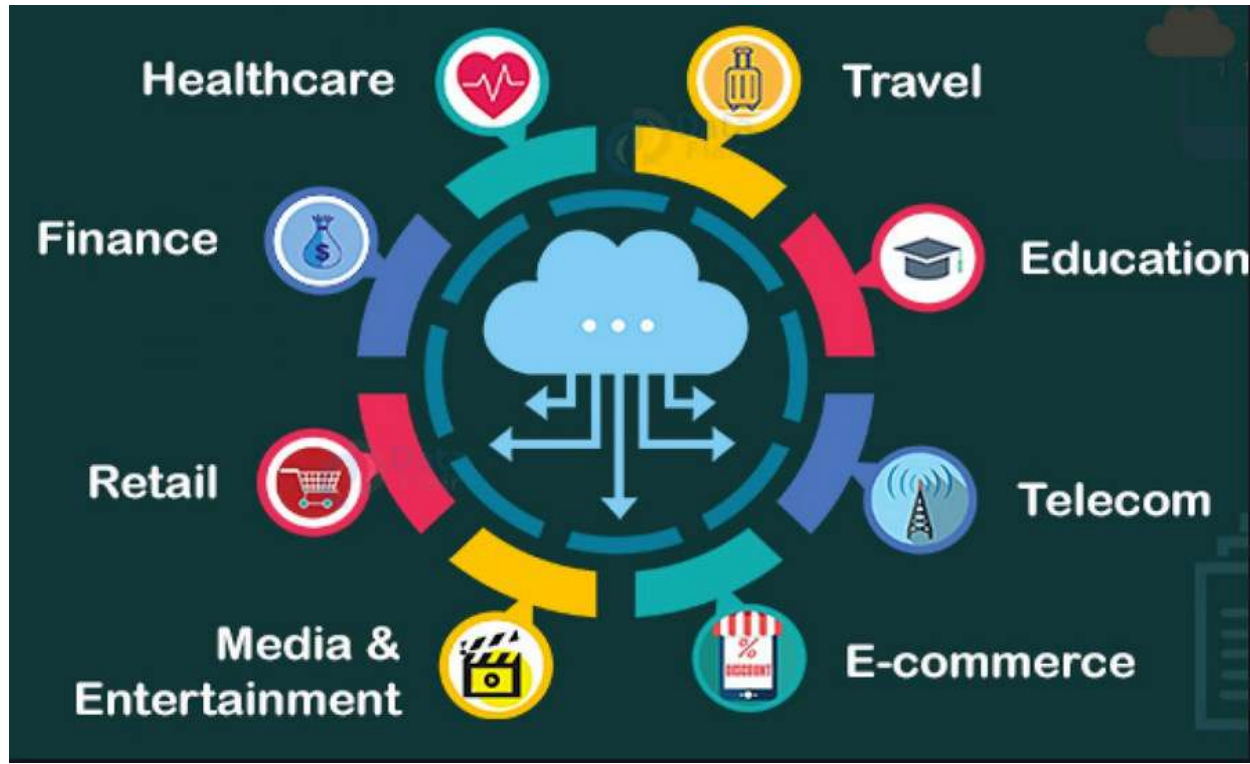
2006's

2010's

2012's

2017's

# Data and computing power



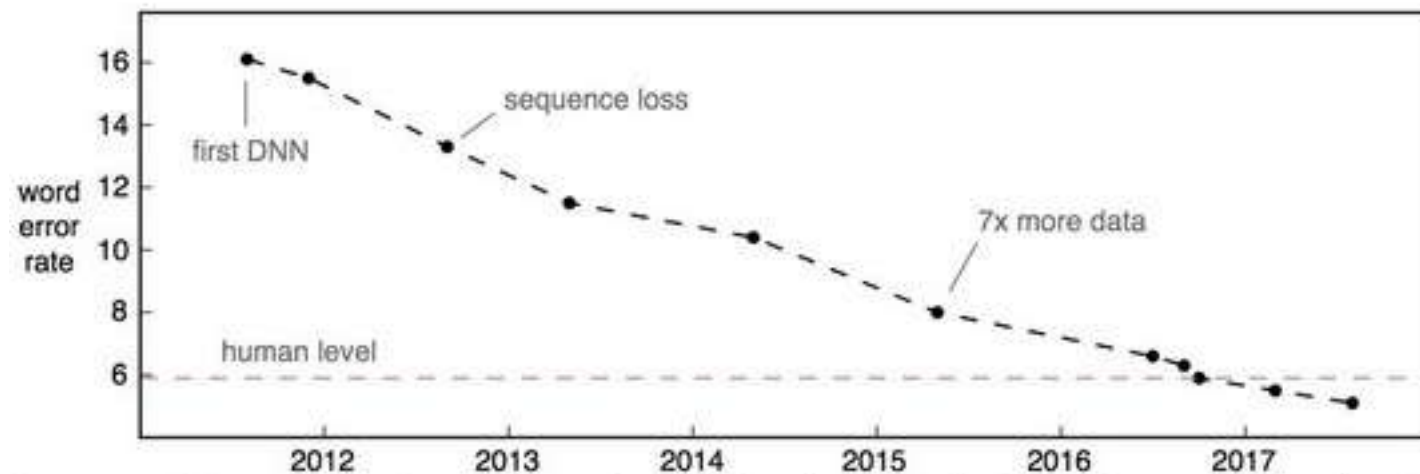
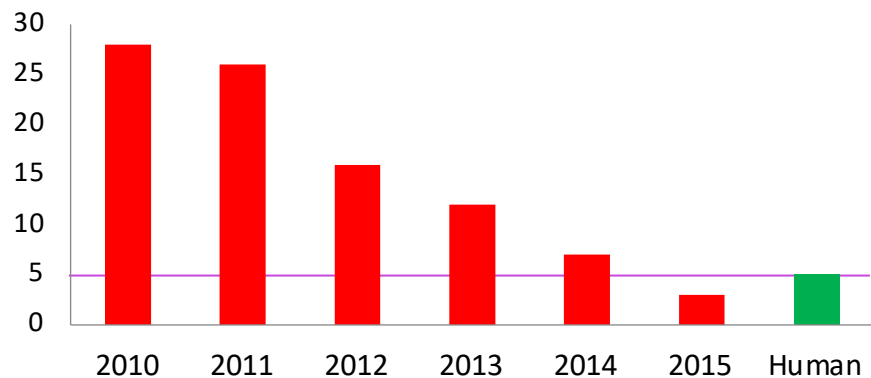
# Image and natural language interpretation



Woman holding a cask of bananas



A group of young people playing fresbee

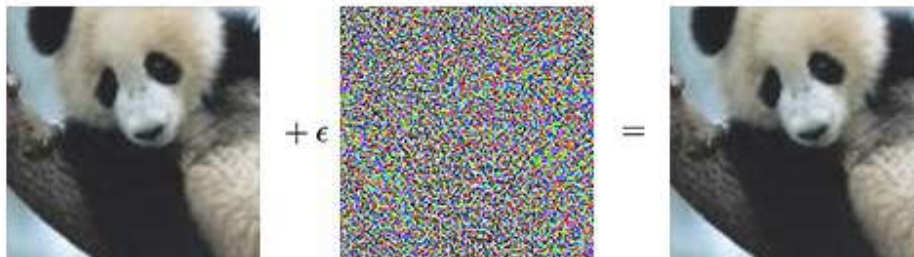


# Some AI applications



- Digital assistants:
  - Home assistants (Alexa)
  - Travel assistants (Waze)
- Driving/travel support:
  - Auto-pilot (Tesla)
  - Ride-sharing apps (Uber, Lyft)
- Customer care:
  - Client service chatbots
- Online recommendations:
  - Friend recommendations (Facebook)
  - Purchase recommendations (Amazon)
  - Movie recommendations (Netflix)
- Media and news:
  - Ad placement (Google)
  - News curation
- Healthcare:
  - Medical image analysis
  - Treatment plan recommendation
- Financial services:
  - Credit risk scoring
  - Loan approval
  - Fraud detection
- Job market:
  - Resume prioritization
- Judicial system:
  - Recidivism prediction (Compass)

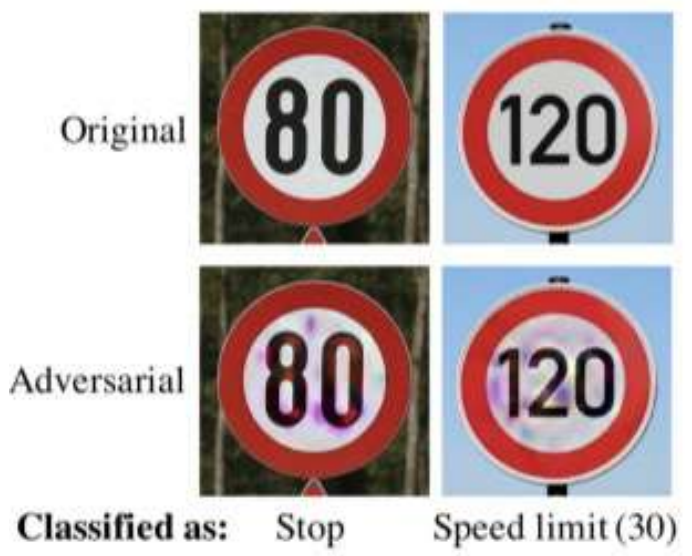




"panda"  
57.7% confidence

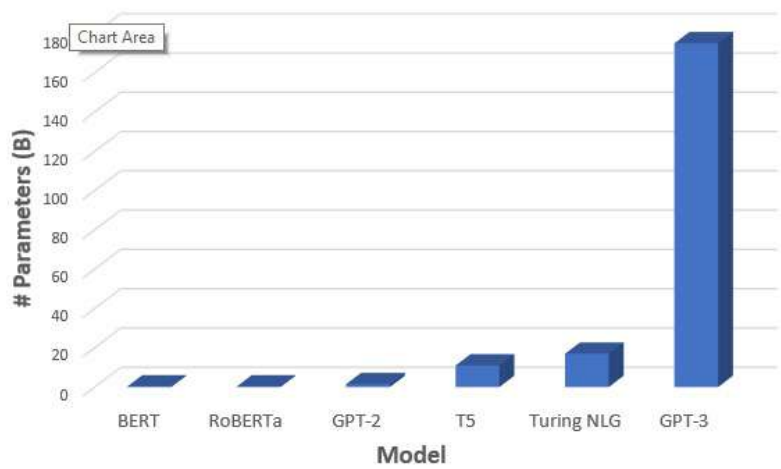
"gibbon"  
99.3% confidence

# AI limitations



Struzzo      Cassaforte      Negozio di scarpe      Aspirapolvere

- Narrow AI
  - Solves well specific problems
- Lack of robustness and adaptability
- Needs a lot of resources
  - Data and computing power



# Ethical issues -- examples

Gender-biased  
Apple credit card  
approval process



Discrimination  
in ride-sharing  
dynamic pricing



Gender-  
biased  
recruitment  
software



IBM Confidential

Chatbot that  
exhibited  
racist speech

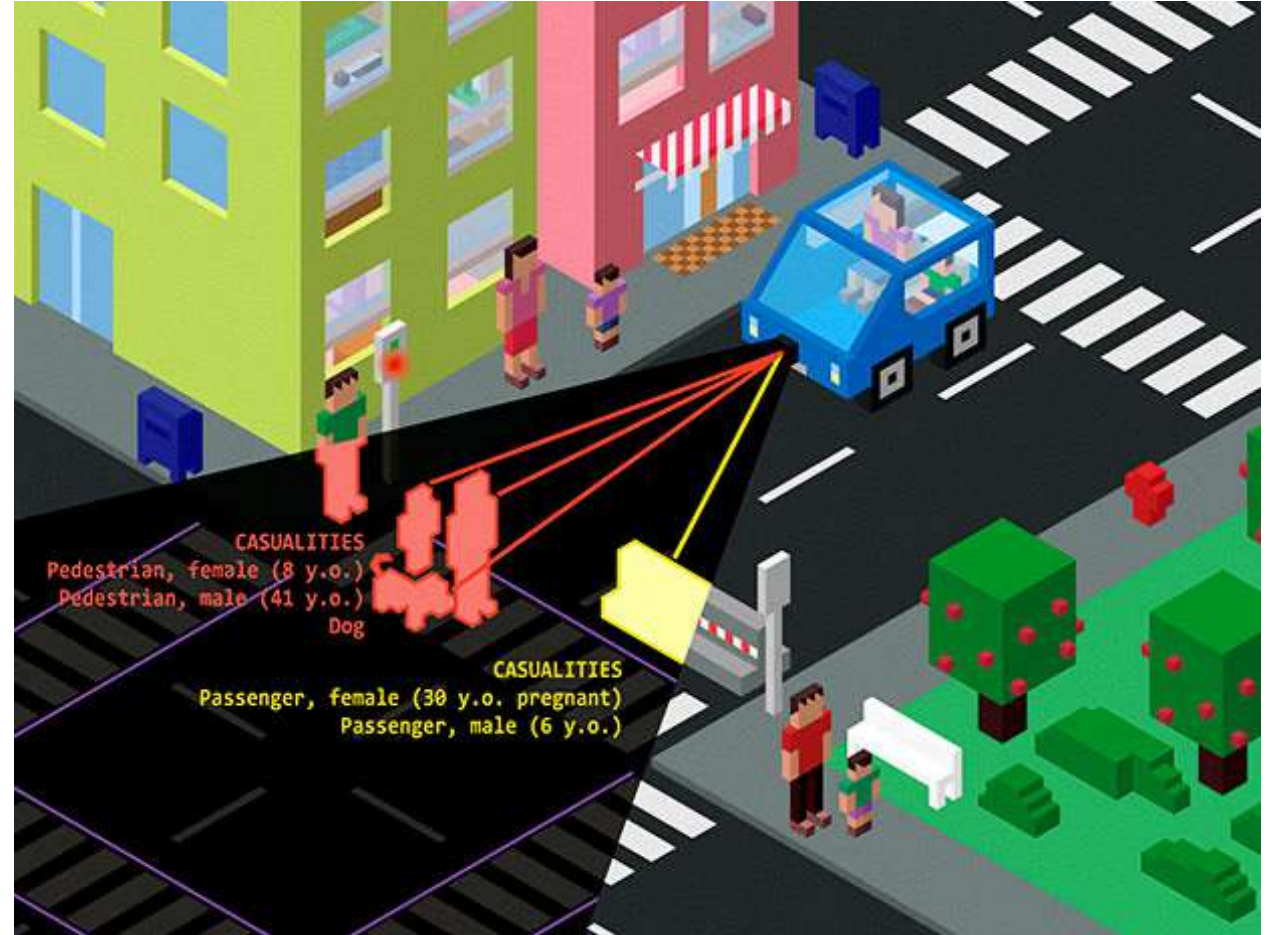


Unethical  
usage  
of personal  
data





Can we trust AI's decisions?



# AI Ethics



Multidisciplinary field of study



How to optimize AI's beneficial impact while reducing risks and adverse outcomes



How to design and build AI systems that are aware of the values and principles to be followed in the deployment scenarios



To identify, study, and propose technical and nontechnical solutions for ethics issues arising from the pervasive use of AI in life and society

# Main AI Ethics issues

## AI needs data

- Data privacy and governance

## AI is often a black box

- Explainability and transparency

## AI can make or recommend decisions

- Fairness and value alignment

## AI is based on statistics and has always a small percentage of error

- Who is accountable if mistakes happen?

## AI can profile people and manipulate their preferences

- Human and moral agency

## AI is very pervasive and dynamic

- Larger negative impacts for tech misuse
- Fast transformation of jobs and society

## Good or bad use of the technology

- Autonomous weapons and mass surveillance
- UN Sustainable Development Goals

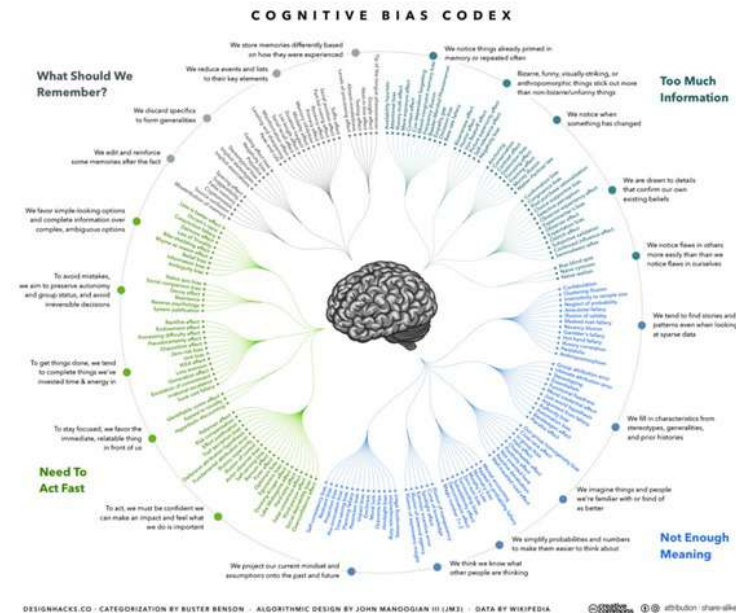
# AI is not a neutral technology

- Misuse must be avoided
- But AI needs to be designed and developed with the right properties
  - Fair, explainable, robust, ...



# AI fairness

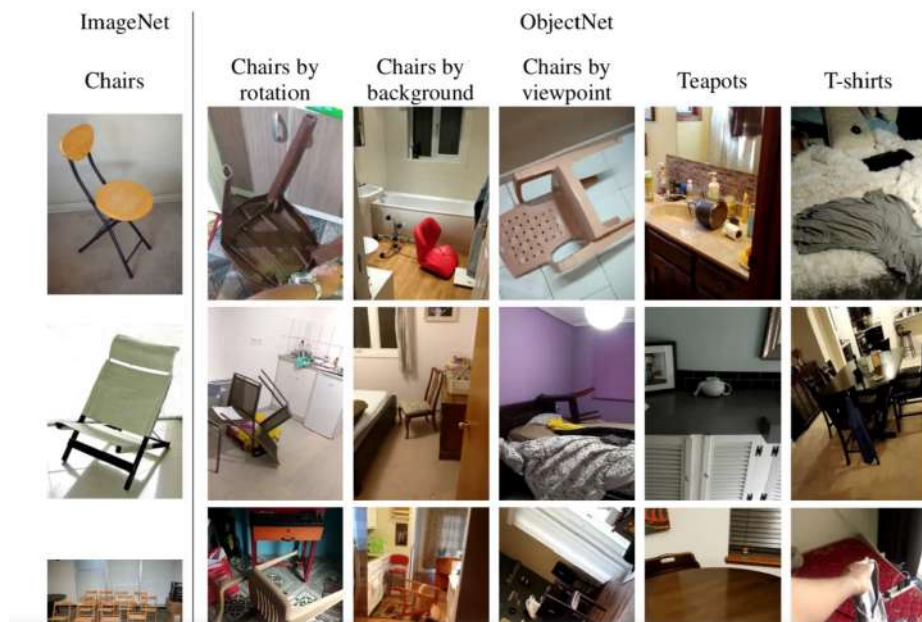
- Bias: prejudice for or against something
- As a consequence of bias, one could behave unfairly to certain groups compared to others
- Why should AI be biased?
  - Trained on data provided by people, and people are biased





# AI bias: ImageNet

- 14M images, used to train image interpretation AI systems
- Bias in the data distribution and in the data labels (Mturk people)



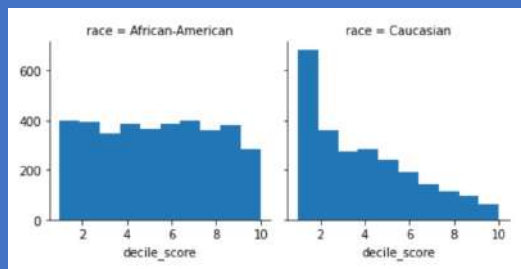
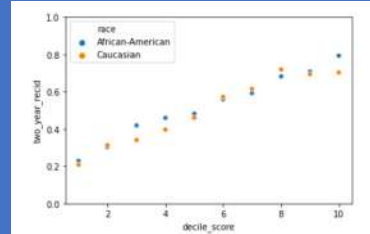
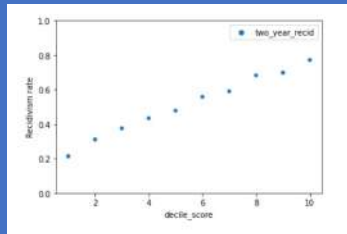


# Mortgage application: bias not just from data



- Training data
  - Ex. : correlation gender-acceptance
- Design decisions:
  - Ex.: prioritized motivations for loan applications
  - Buying a house
  - Paying school fees
  - Paying legal fees
    - Loan applications with these motivations are prioritized
    - If one of them is omitted, the relevant community will be penalized

# AI bias: which is the correct definition of fairness?



- Overall accuracy is the same, regardless of race (**overall accuracy equality**)
- Likelihood of recidivism among defendants labeled as medium or high risk is similar, regardless of race (**predictive parity**)
- But ... false positive and false negative rates are very different

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

# Many decision points

- **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

Source: *Fairness and Machine Learning* by Solon Barocas, Moritz Hardt, Arvind Narayanan (<https://www.fairmlbook.org>)

Name	Closest relative	Note	Reference
Statistical parity	Independence	Equivalent	Dwork et al. (2011)
Group fairness	Independence	Equivalent	
Demographic parity	Independence	Equivalent	
Conditional statistical parity	Independence	Relaxation	Corbett-Davies et al. (2017)
Darlington criterion (4)	Independence	Equivalent	Darlington (1971)
Equal opportunity	Separation	Relaxation	Hardt, Price, Srebro (2016)
Equalized odds	Separation	Equivalent	Hardt, Price, Srebro (2016)
Conditional procedure accuracy	Separation	Equivalent	Berk et al. (2017)
Avoiding disparate mistreatment	Separation	Equivalent	Zafar et al. (2017)
Balance for the negative class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Balance for the positive class	Separation	Relaxation	Kleinberg, Mullainathan, Raghavan (2016)
Predictive equality	Separation	Relaxation	Chouldechova (2016)
Equalized correlations	Separation	Relaxation	Woodworth (2017)
Darlington criterion (3)	Separation	Relaxation	Darlington (1971)
Clearly model	Sufficiency	Equivalent	Clearly (1966)
Conditional use accuracy	Sufficiency	Equivalent	Berk et al. (2017)
Predictive parity	Sufficiency	Relaxation	Chouldechova (2016)
Calibration within groups	Sufficiency	Equivalent	Chouldechova (2016)
Darlington criterion (1), (2)	Sufficiency	Relaxation	Darlington (1971)

AI  
explainability:  
AI systems  
cannot be  
black boxes

The **General Data Protection Regulation (GDPR)**

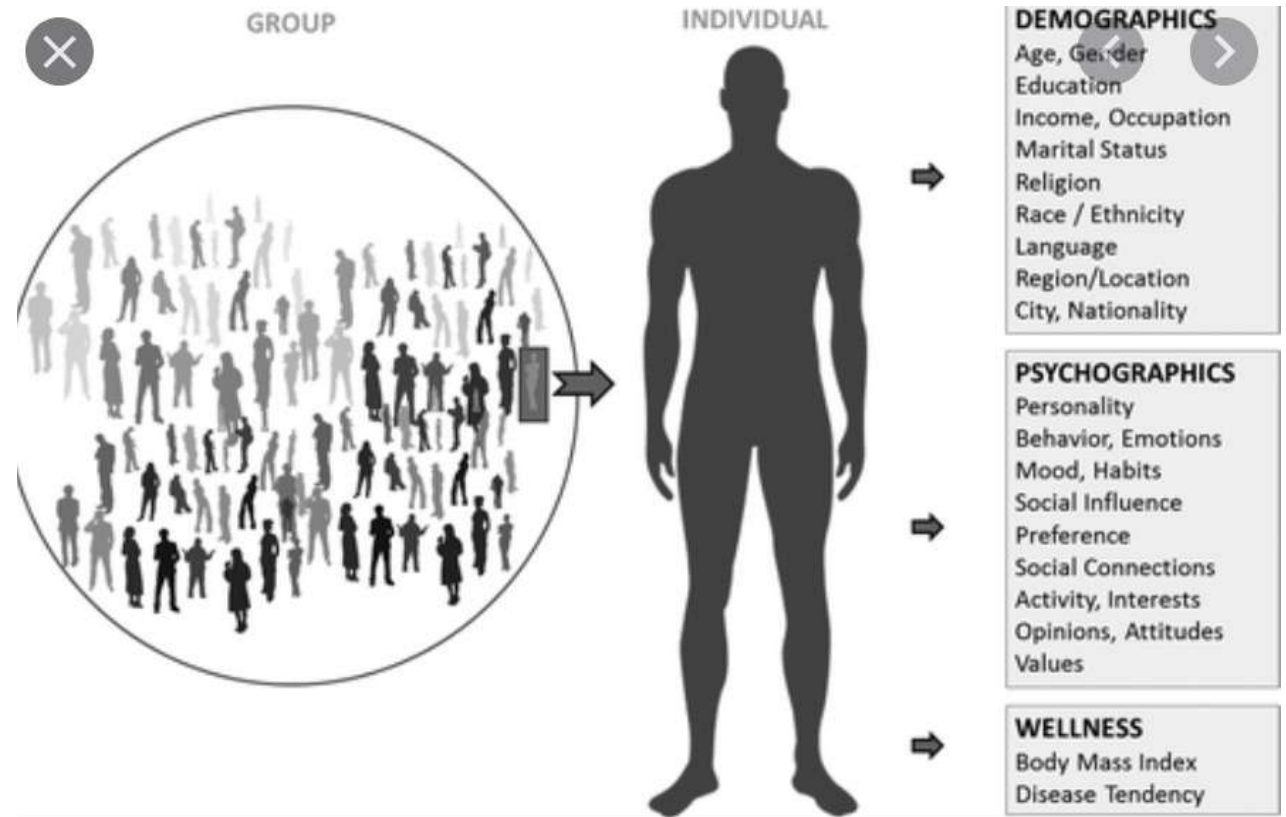
- Limits to **decision-making** based solely on **automated processing** and profiling (Art.22)
- Right to be provided with **meaningful information** about the **logic** involved in the decision ( Art.13 (2) f. and 15 (1) h)

# Data handling: the General Data Protection Regulation (GDPR)



# Profiling and manipulation

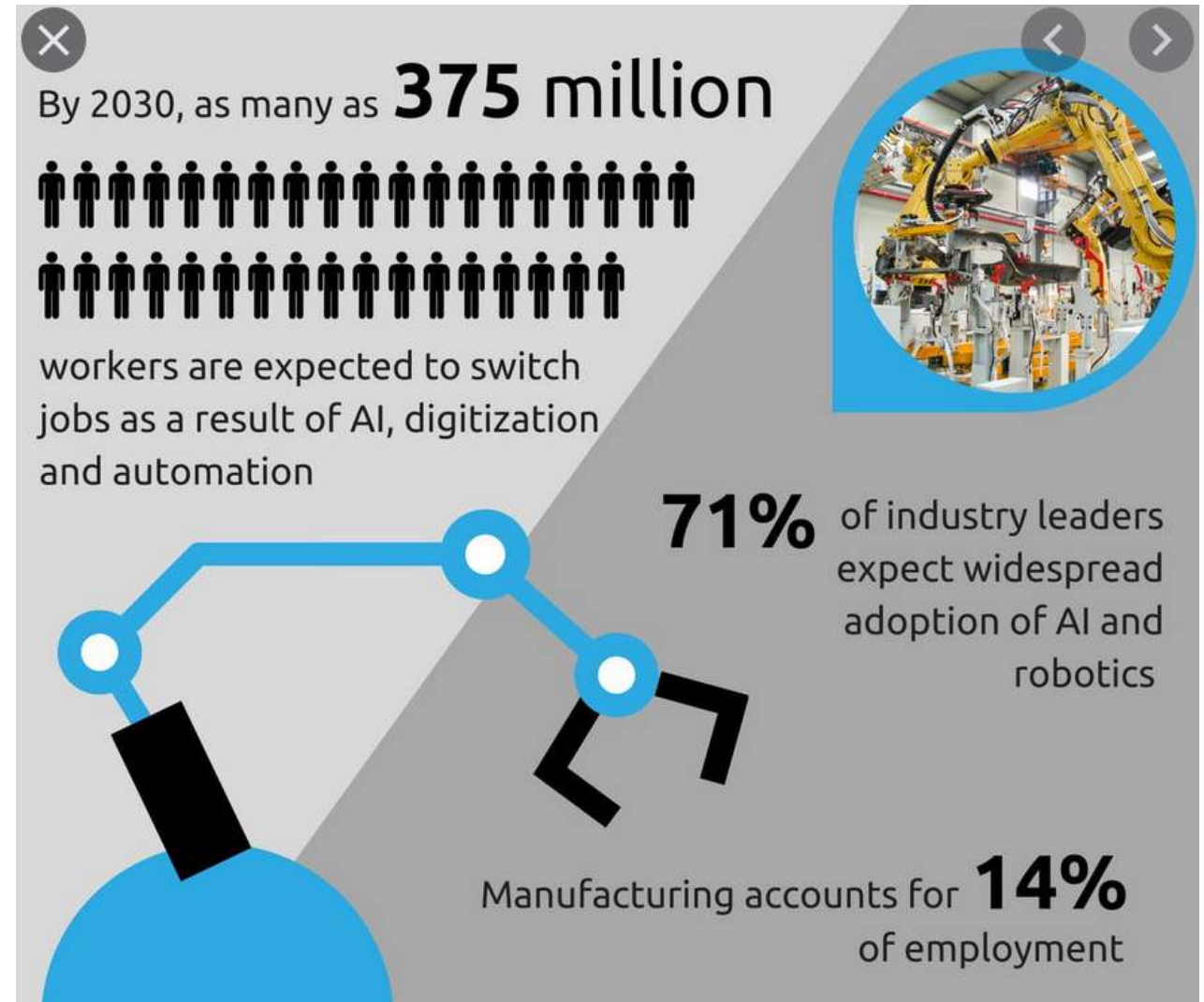
- From actions to profiles
  - Like, text, images, follow, ...
- AI can infer our preferences, and use them to advertise products that we probably like
  - Easier if our preferences are bipolar





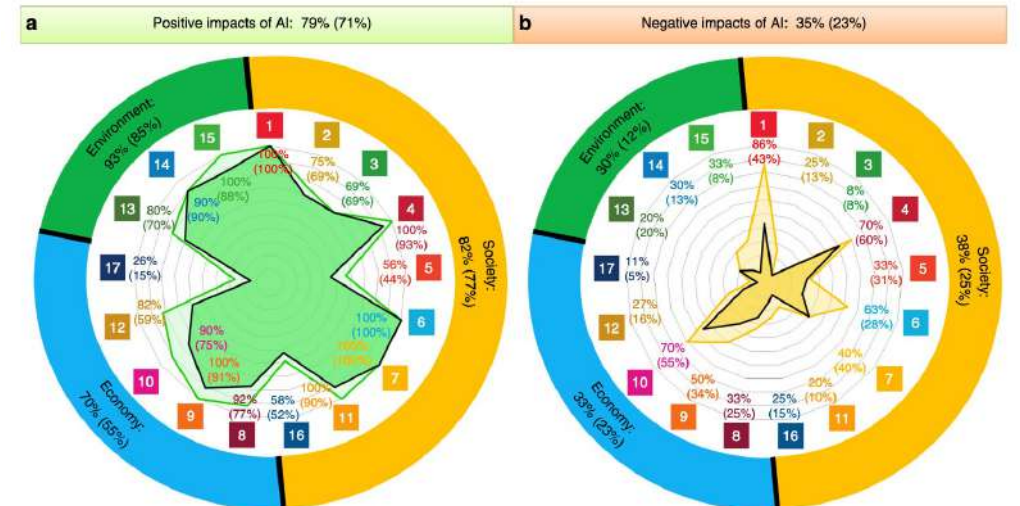
# Impact on the workforce

- Many jobs will disappear, and many others will be created
- All jobs will change



# A vision of the future (2030)

- 17 goals, 169 targets
- Very difficult path
  - The pandemic has worsened the situation
- AI can help in achieving the SDGs
- COVID: vaccines in less than one year!



# IBM, technology, and AI

- 110 years
- Hardware e software
- Enterprise AI: AI solutions for other companies
  - Banks and financial institutions
  - Governments
  - Aeroports
  - Hospitals
  - ...



Summit, IBM



Quantum computer, IBM



Chess: IBM Deep Blue, 1997



Jeopardy: IBM Watson, 2011



Project Debater, 2020

# IBM Principles of Trust and Transparency (2017)



The purpose of AI is to **augment** human intelligence



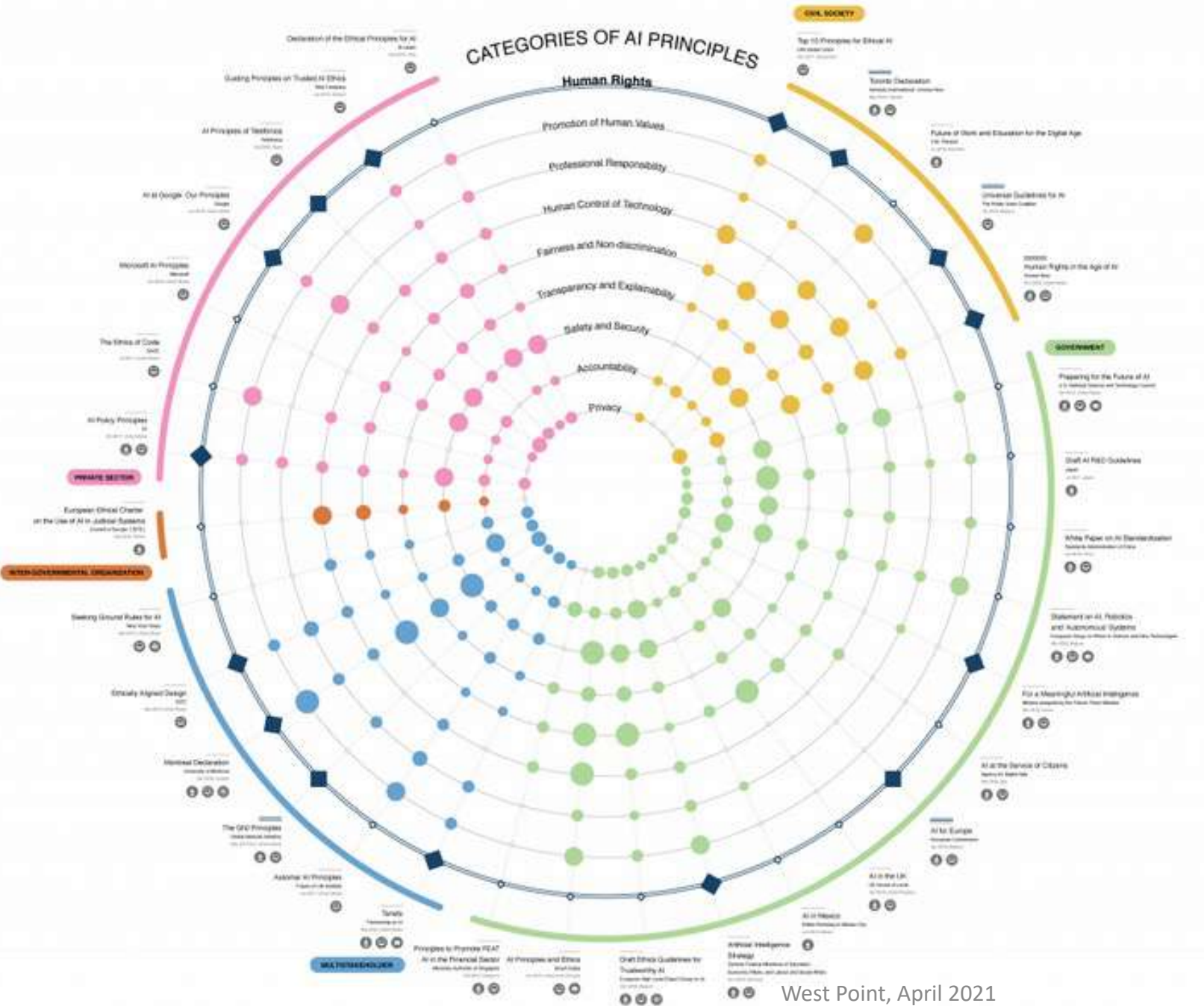
**Data** and insights belong to their creator



New technology, including AI systems, must be **transparent** and **explainable**



# AI PRINCIPLES in the world – a comprehensive view



## Actors:

- Private sector
- Inter-governmental
- Multistakeholder
- Governments
- Civil society

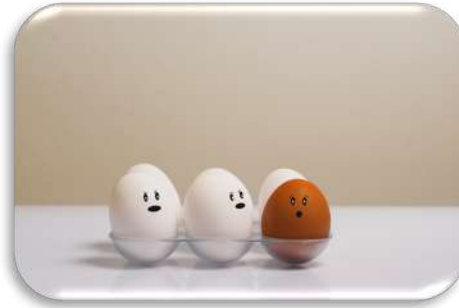
## Main themes:

- Human rights
- Human values
- Responsibility
- Human control
- Fairness
- Transparency and explainability
- Safety and Security
- Accountability
- Privacy

Principled AI Project,  
 Berkman Klein's Cyberlaw  
 Clinic, 2019



# What does it mean to TRUST a decision made by a machine? (Other than it is accurate and respect privacy)



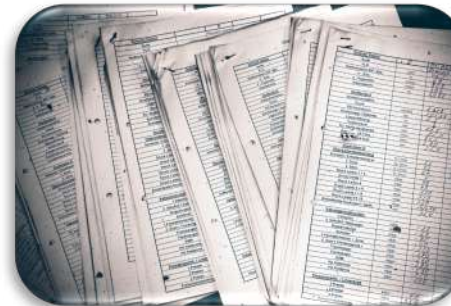
Is it **fair**, or is it going to make discriminatory decisions?



Is it possible to understand **why** it made that decision, or is it a black box?



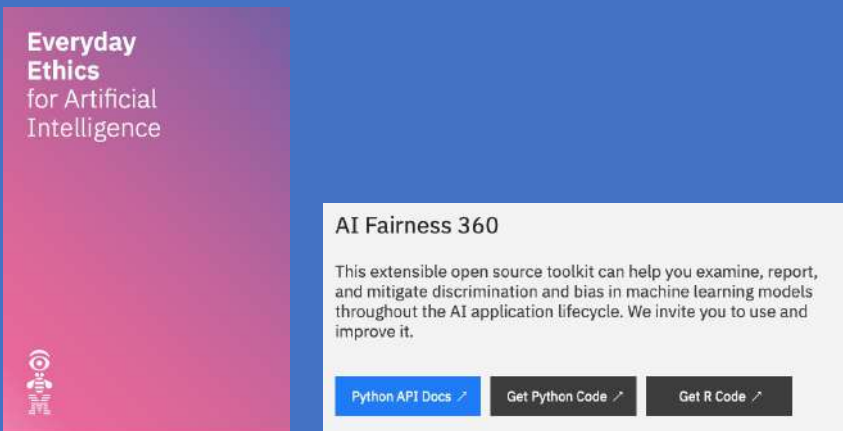
Is it **robust**?



Is it **transparent**?



# AI fairness at IBM



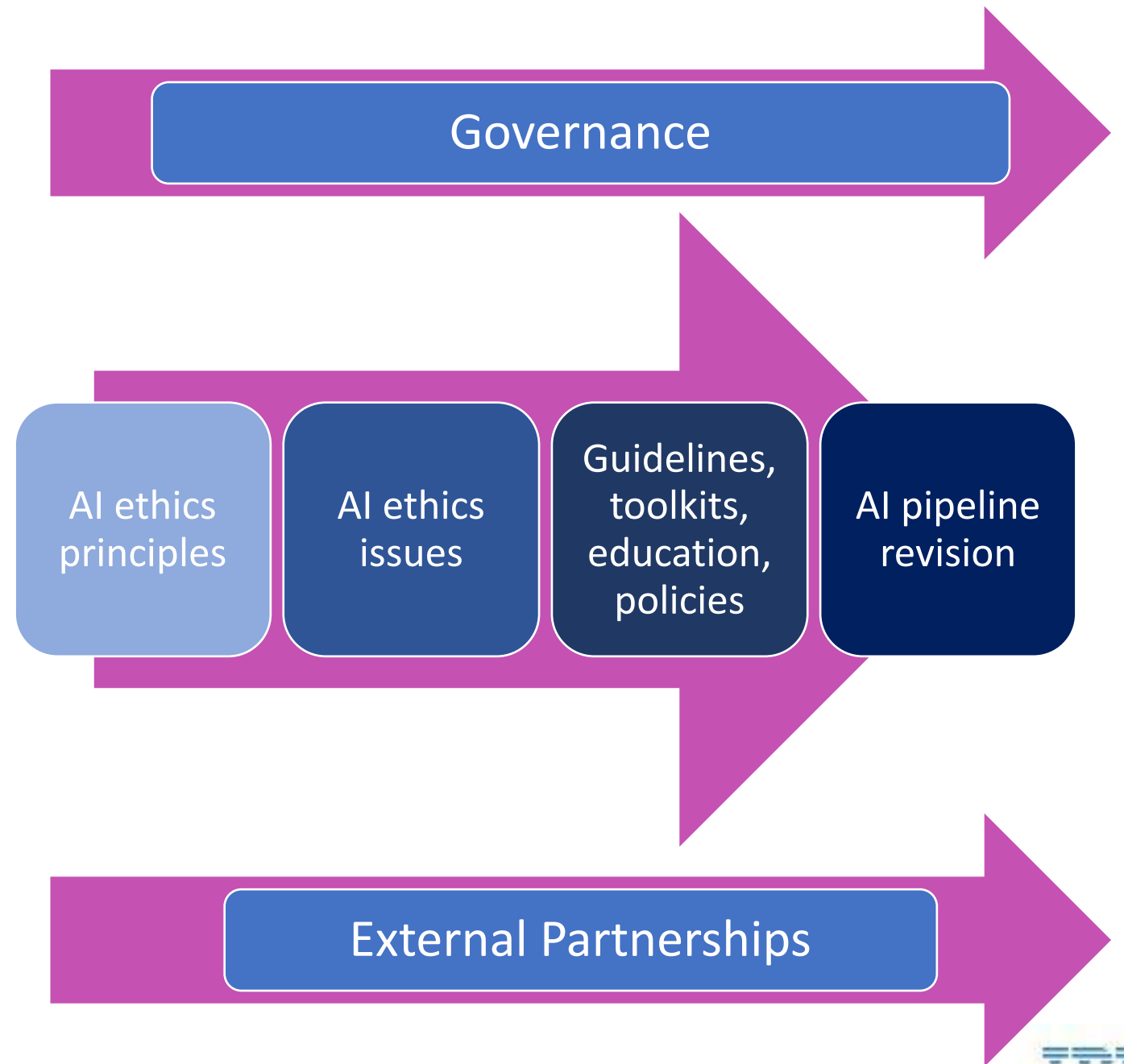
- Technical solutions to detect and mitigate AI bias
  - Research work
  - Watson OpenScale
  - Open-source libraries: AI fairness 360
- Developers' education and training
  - AI bias education modules for all IBMers
  - Developers' awareness material
  - Revised methodologies for the AI pipeline
  -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

# AI transparency at IBM

- What is the **intended use** of the service output?
- What **algorithms** or techniques does this service implement?
- Which datasets was the service **tested** on?
- Describe the **testing methodology** and **test results**.
- Are you aware of possible examples of **bias**, **ethical** issues, or other **safety risks** as a result of using the service?
- Are the service outputs **explainable** and/or interpretable?
- For each dataset used by the service:
  - Was the dataset checked for **bias**?
  - What efforts were made to ensure that it is **fair** and **representative**?
  - Does the service implement and perform any **bias detection and remediation**?
- What is the **expected performance** on unseen data or data with different distributions?
- Was the service checked for **robustness against adversarial attacks**?
- When were the models last updated?

- 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

# From principles to practice: a multi-dimensional space



# Governance: the IBM AI Ethics board

- Mission
  - Awareness and coordination
  - Internal education and retraining
  - Linking research to services and platforms
  - Advice to business units
  - Internal governance framework
  - Define policies and advice regulators
- Risk-based approach for the BUs
  - Vetting based on three dimensions (tech, use, client)



# Partnerships

Academia  
Companies  
Governments  
Civil society  
organizations

Multi-disciplinary and  
multi-stakeholder

## Asilomar AI principles

The infographic lists 23 principles under three categories: RESEARCH, ETHICS AND VALUES, and LONGER-TERM ISSUES. It includes an illustration of a robot head and the Future of Life Institute logo.

- RESEARCH**
  1. Research goal
  2. Research funding
  3. Science-policy link
  4. Research culture
  5. Race avoidance
- ETHICS AND VALUES**
  6. Safety
  7. Failure transparency
  8. Judicial transparency
  9. Responsibility
  10. Value alignment
  11. Human values
  12. Personal privacy
  13. Liberty and privacy
  14. Shared benefit
  15. Shared prosperity
  16. Human control
  17. Non-subversion
  18. AI arms race
- LONGER-TERM ISSUES**
  19. Capability caution
  20. Importance
  21. Risks
  22. Recursive self-improvement
  23. Common good



AAAI / ACM conference on  
**ARTIFICIAL INTELLIGENCE,  
ETHICS, AND SOCIETY**



**AI for Good  
Global Summit**  
An ITU experience

Version II - For Public Discussion  
IEEE  
Advancing Technology  
for Humanity  
**ETHICALLY  
ALIGNED DESIGN**  
A Vision for Prioritizing Human Well-being  
with Autonomous and Intelligent Systems

Partnership on AI  
to benefit people and society

One organization

to develop and share the best practices for using and developing AI technologies and providing a global platform to discuss how AI will influence people and society.

7 Thematic Pillars

- Safety Critical AI
- Fair, Transparent, and Accountable AI
- AI, Labour and the Economy
- Collaborations between People and AI systems
- AI and Social Good
- Social and Societal Influences of AI
- Special Initiatives

WORLD  
ECONOMIC  
FORUM

INDEPENDENT  
HIGH-LEVEL EXPERT GROUP ON  
ARTIFICIAL INTELLIGENCE  
SET UP BY THE EUROPEAN COMMISSION

**ETHICS GUIDELINES  
FOR TRUSTWORTHY AI**

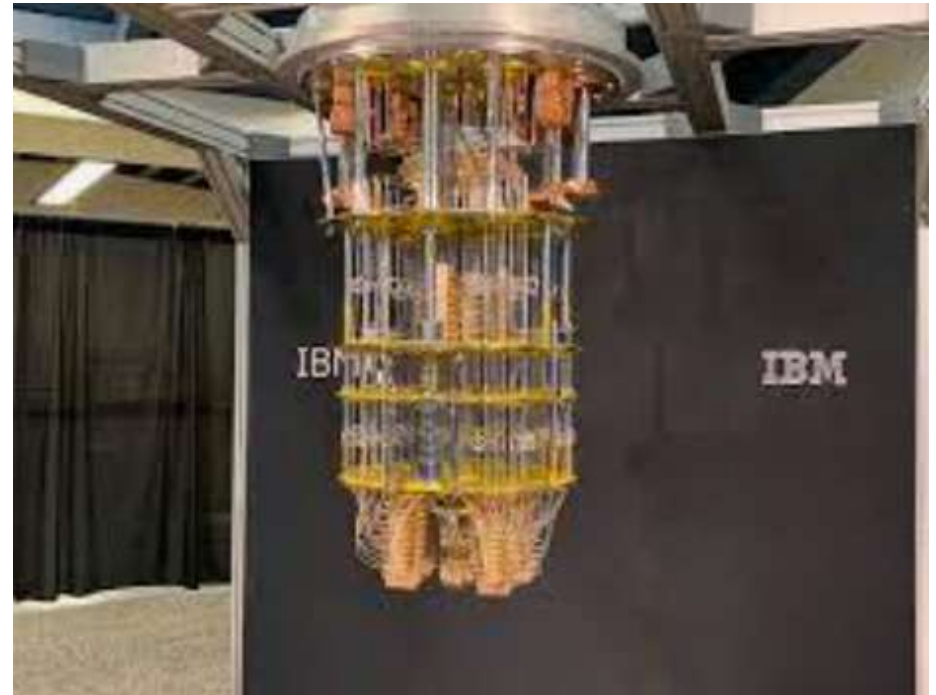
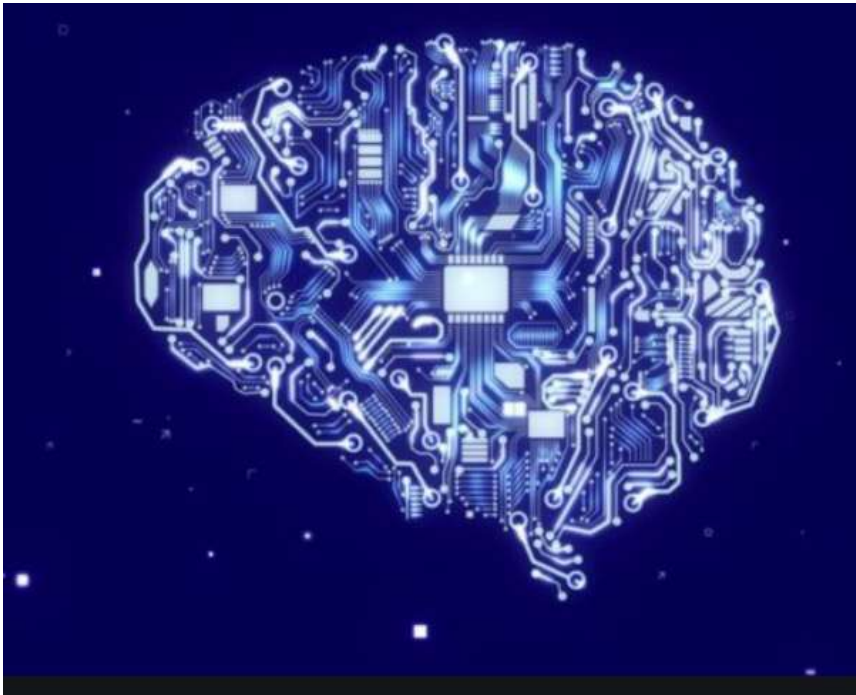
INDEPENDENT  
HIGH-LEVEL EXPERT GROUP ON  
ARTIFICIAL INTELLIGENCE  
SET UP BY THE EUROPEAN COMMISSION

**POLICY AND INVESTMENT RECOMMENDATIONS  
FOR TRUSTWORTHY AI**



# Not just AI

- 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?

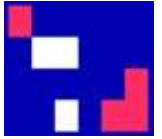


# Useful links

- IBM Approach to AI Ethics:
  - External website: <https://www.ibm.com/artificial-intelligence/ethics>
  - Trusted AI for business: <https://www.ibm.com/watson/ai-ethics/>
- Educational material:
  - Everyday Ethics for AI: <https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf>
- External articles:
  - Harvard Business Review article, 2020: <https://hbr.org/2020/11/how-ibm-is-working-toward-a-fairer-ai>
- Global studies:
  - IBM IBV study on “Advancing AI ethics beyond compliance”:  
<https://www.ibm.com/thought-leadership/institute-business-value/report/ai-ethics>
- Public policies:
  - IBM Policy Lab: <https://www.ibm.com/policy/>
  - AI precision regulation: <https://www.ibm.com/blogs/policy/ai-precision-regulation/>
  - Facial recognition: <https://www.ibm.com/blogs/policy/facial-recognition/>
  - Response to COVID-19: <https://www.ibm.com/thought-leadership/covid19/>
- Open-source toolkits:
  - AI fairness 360: <https://aif360.mybluemix.net/>
  - AI explainability 360: <https://aix360.mybluemix.net/>
  - AI factsheet 360: <http://aifs360.mybluemix.net/>

Thank you!





# Fairness in algorithmic decision making

Francesca Lagioia

Giovanni Sartor

European University Institute



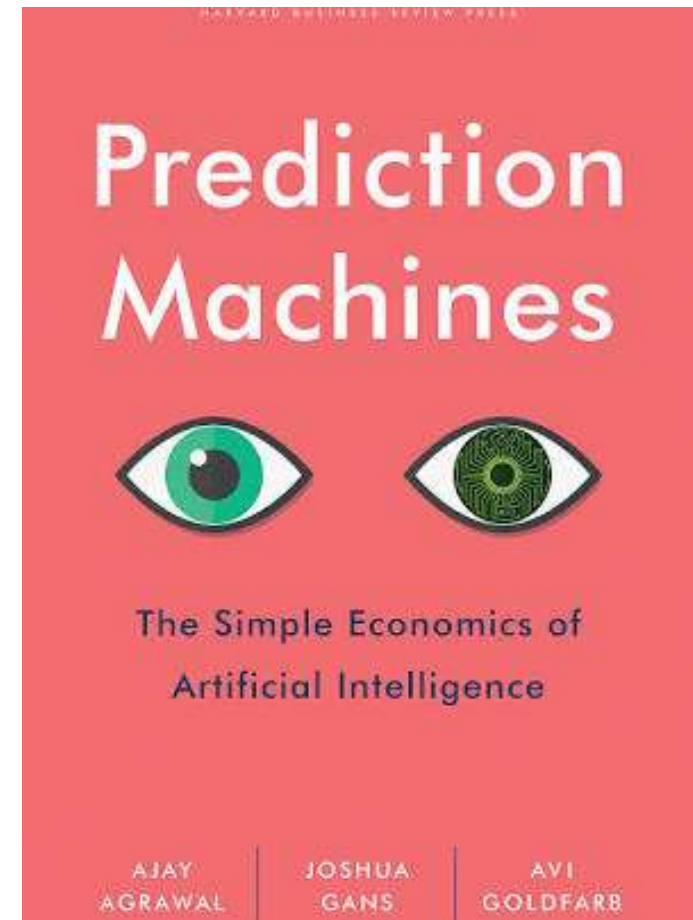
# Outline

- AI in decision making concerning individuals
  - Possible causes of unfairness
- The principle of Fairness and its substantive dimension
- AI unfairness
  - The COMPAS predictive system and the Loomis case
  - A toy example and the criteria for assessing fairness

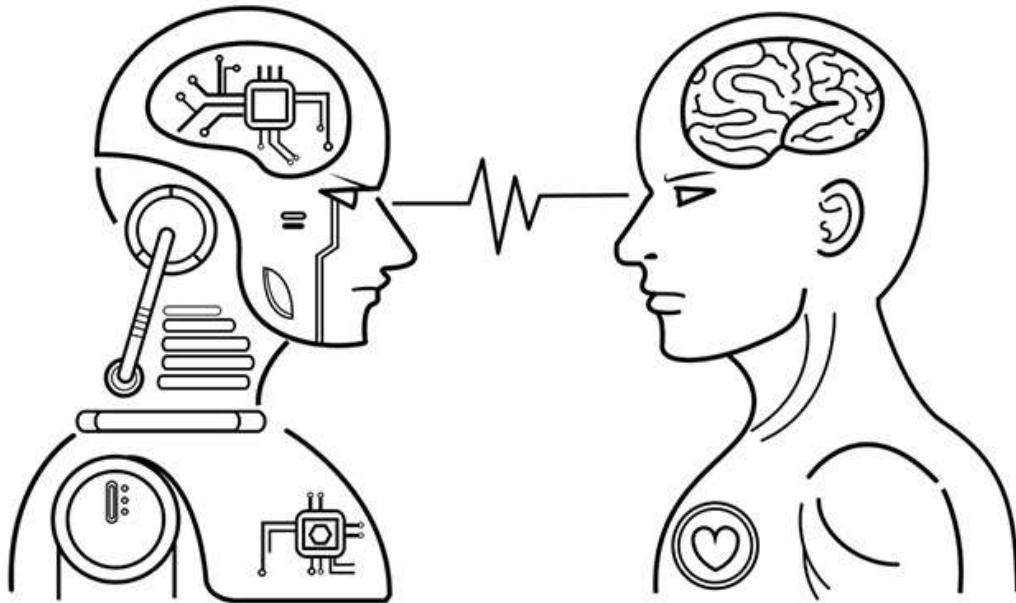


# AI in decision making concerning individuals: fairness and discrimination

- The combination of AI and Big Data enables automated decision-making even in domains that require complex choices, based on multiple factors, and on non-predefined criteria.
- In recent years, a wide debate has taken place on prospects and risks of algorithmic assessments and decisions concerning individuals



# Are AI systems better than humans in assessing us?



In many domains automated predictions and decisions are not only **cheaper**, but also **more precise and impartial** than human ones.

- AI can **avoid typical fallacies of human psychology** (overconfidence, loss aversion, anchoring, confirmation bias, representativeness heuristics, etc.), and the widespread human **inability to process statistical data**, as well as **typical human prejudice** (concerning, e.g., ethnicity, gender, or social background).
- In many assessments and decisions —on investments, recruitment, creditworthiness, or also on judicial matters, such as bail, parole, and recidivism— algorithmic systems have **often performed better**, according to usual standards, than human experts.

## Or not?

Others have underscored the possibility that algorithmic decisions may be **mistaken** or **discriminatory**.

- Only in rare cases will algorithms engage in explicit unlawful discrimination, so-called **disparate treatment**, basing their outcomes on prohibited features (predictors) such as race, ethnicity or gender.
- More often a system's outcome will be discriminatory due to its **disparate impact**, i.e., since it disproportionately affects certain groups, without an acceptable rationale



# Systems reproducing the strengths and weaknesses of humans in making judgments

---



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

- For example, a recruitment system trained on the past hiring decisions will learn to emulate the managers' assessment of the suitability of candidates, rather than to directly predict an applicant's performance at work. If past decisions were influenced by prejudice, the system will reproduce the same logic.

# Prejudice in the training set

---

Prejudice baked into training sets may persist even if the inputs (the predictors) to automated systems do not include forbidden discriminatory features (e.g. ethnicity or gender.)

This may happen whenever a **correlation exists between discriminatory features and some predictors**

- Assume, for instance, that a prejudiced human resources manager did not hire applicants from a certain ethnic background, and that people with that background mostly live in certain neighbourhoods. A training set of decisions by that manager will teach the systems not to select people from those neighbourhoods, which would entail continuing to reject applications from the discriminated-against ethnicity. (Kleinberg et al (2019)).



# Systems biased against groups

---

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

- Assume, for instance, that the **future performance** of employees (the target of interest in job hiring) is only measured by the **number of hours worked in the office**. This outcome criterion will lead to past hiring of women —who usually work for fewer hours than men, having to cope with family burdens— being considered less successful than the hiring of men; based on this correlation (as measured on the basis of the biased proxy), the systems will predict a poorer performance of female applicants.





# System's biases embedded in the predictors

In other cases, mistakes and discriminations may pertain to the machine-learning system's biases embedded in the predictors.

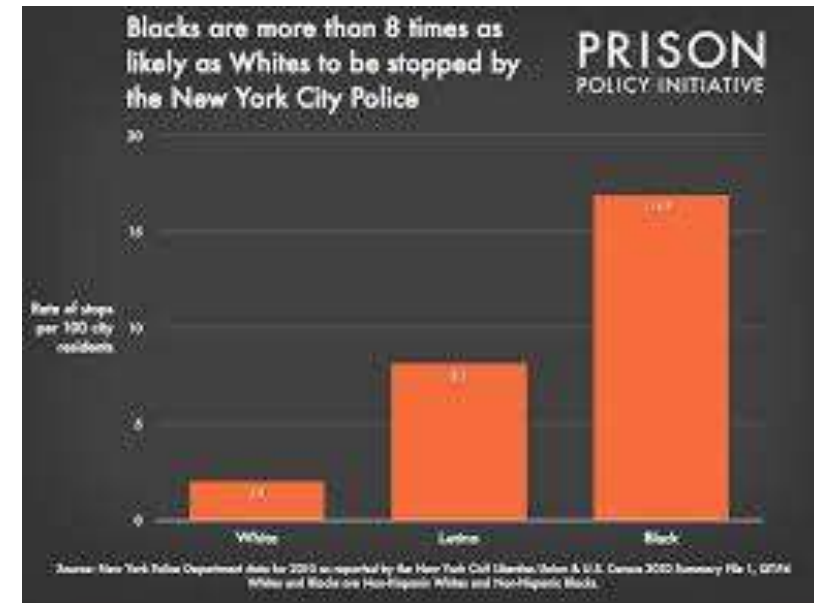
A system may perform unfairly, since it uses a favourable predictor (input feature) that only applies to members of a certain group (e.g., the fact of having attended a socially selective high-education institution).

Unfairness may also result from taking biased human judgements as predictors (e.g., recommendation letters).

# Data set that does NOT reflect the statistical composition of the population

Finally, unfairness may derive from a data set that does reflect the statistical composition of the population.

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



- Members of a certain group may also suffer prejudice when that group is only represented by a very small subset of the training set,
- This will reduce the accuracy of predictions for that group (e.g., consider the case of a firm that has appointed few women in the past and which uses its records of past hiring as its training set).



# Challenging the unfairness of automated decision- making

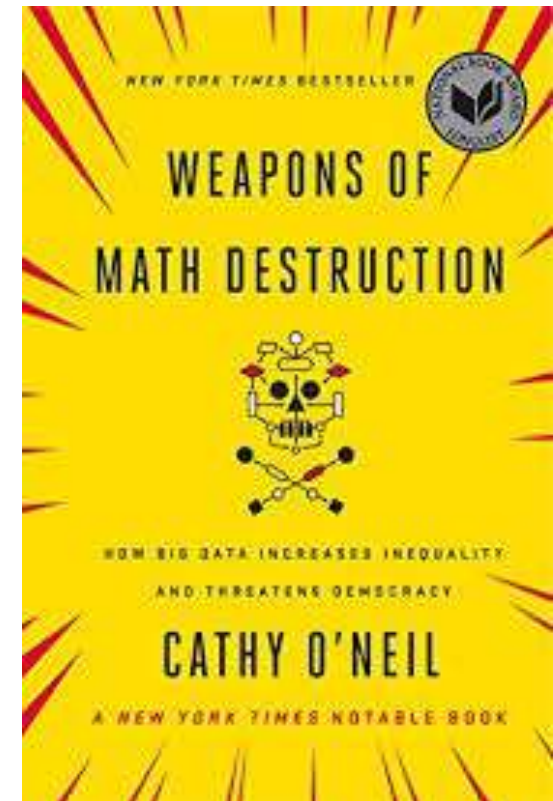
It has been observed that it is difficult to challenge the unfairness of automated decision-making.

Challenges raised by the individuals concerned, even when justified, may be disregarded or rejected because they interfere with the system's operation, giving rise to additional **costs and uncertainties**.

In fact, predictions of machine-learning systems are based on **statistical correlations**, against which it may be **difficult to argue** on the basis of individual circumstances, even when exceptions would be justified.

# Weapons of math destruction

“An algorithm processes a slew of statistics and comes up with a probability that a certain person might be a bad hire, a risky borrower, a terrorist, or a miserable teacher. That probability is distilled into a score, which can turn someone’s life upside down. And yet when the person fights back, “suggestive” countervailing evidence simply won’t cut it. The case must be ironclad. The human victims of WMDs, we’ll see time and again, are held to a far higher standard of evidence than the algorithms themselves”. (O’Neil (2016))



# Or not?

---

[W]ith appropriate requirements in place, the use of algorithms will make it possible to more easily examine and interrogate the entire decision process, thereby making it far easier to know whether discrimination has occurred. By forcing a new level of specificity, the use of algorithms also highlights, and makes transparent, central **trade-offs among competing values**. Algorithms are not only a threat to be regulated; with the right safeguards in place, they have **the potential to be a positive force for equity**

(Kleinberg, Ludwig, Mullainathan, e Sunstein (2018, 113)).





# Challenging the unfairness of automated decision-making

These criticisms have been countered by observing that **algorithmic systems**, even when based on machine learning, are **more controllable** than human decision-makers, their **faults can be identified** with precision, and **they can be improved and engineered** to prevent unfair outcomes.



# Should we exclude the use of automated decision-making?

---

It seems that issues that have just been presented should not lead us to exclude categorically the use of automated decision-making.

The alternative to automated decision-making is not perfect decisions but human decisions with all their flaws: a biased algorithmic system can still be fairer than an even more biased human decision-maker.



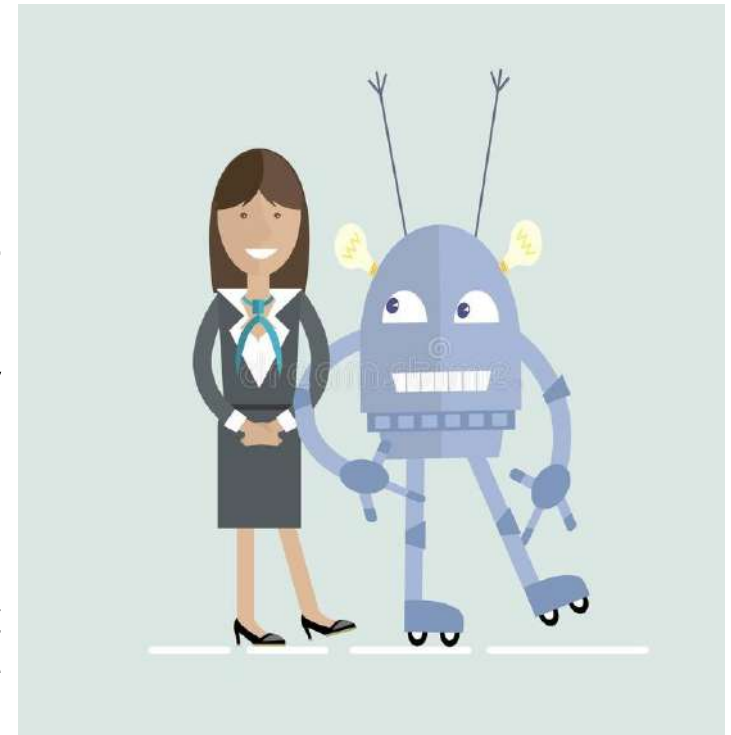
# Humans + Algorithms?

---

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

In fact, AI systems have demonstrated an ability to successfully also act in domains traditionally entrusted the trained intuition and analysis of humans, e.g., medical diagnosis, financial investment, granting of loans, etc.

The future challenge will consist in finding the best combination between human and AI, taking into account the capacities and the limitations of both.



# Substantive Fairness and AI

- Equal and just distribution of **benefits** and **costs**
- Individuals and groups free from unfair **bias**, **discrimination** and **stigmatisation**
- AI decision making: informational fairness + content fairness of inferences/decision  
(avoid prejudice, discrimination, etc.)
  - appropriate mathematical or statistical procedures for profiling,
  - technical and organisational measures to ensure correctness of personal data
  - secure personal data (potential risks, discriminatory effects, etc.)

# The COMPAS system: AI and unfairness

- An actuarial risk assessment instrument to determine:
  - Risk of recidivism and appropriate correctional treatment
- Based on statistical algorithms
- Offenders are classified in three categories: high, medium, low risk
  - Multiple-choice test (137 questions)
  - Static risk variables (e.g., prior criminal history, education, etc.)
  - Dynamic risk variables (e.g., drug abuse, employment status)



# The Loomis case

- In 2013 E. Loomis was charged with driving a stolen vehicle and fleeing from police
- The Districtal Court ordered a presentencing investigation that included the COMPAS risk assessment
- Loomis was classified at high risk for recidivism and sentenced to 6 years imprisonment
- The decision was appealed by Loomis for violation of due process rights (e.g., basic rights of defence):

- COMPAS functioning is unknown

- Its validity can not be verified

- It discriminates on gender and race

- Statistical-based predictions violate the right to individualized decision.



# The Loomis case

In 2016 the Supreme Court of Wisconsin rejected all defendant's arguments.

According to the Supreme Court:

- Statistical algorithms does not violate the right to individualized decisions
- They should be used to “enhance a judge's evaluation of other evidence in the formulation of an individualized sentencing
- Prohibition to base decisions solely on risk scores + obligation to motivate as safeguards of the defendant' rights.
- Considering gender is necessary to achieve statistical accuracy.
- Judges should be informed on the debate concerning COMPAS race discrimination

# The challenges

In 2016 ProPublica published a study (Larson et al. 2016):

**Sample:** 11,757 defendants assessed by COMPAS (2013-2014)

**Objective:** evaluate COMPAS accuracy and fairness

**Methodology:** Comparison between predicted recidivism rates and the rate that actually occurred over 2-year period.

# The challenges

## ProPublica Results:

- Moderate-Low Predictive accuracy (61.2%)
- Black defendants were predicted at a higher risk than they actually were. Probability of high-risk misclassification (45% blacks vs. 23% whites)
- White defendants were often predicted to be less risky than they were. Probability of low-risk misclassification (48% whites vs. 28% blacks).

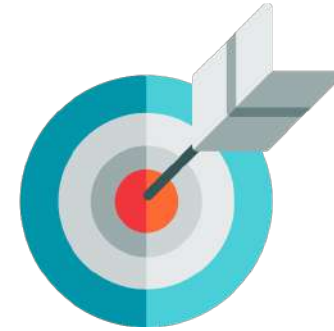
# The rebuttals

According to Northpoint (Dieterich et al 2016) ProPublica made several statistical and technical errors

- The accuracy of COMPAS predictions > accuracy of human judgments
- General Recidivism Risk Scale is equally accurate for blacks and whites
- COMPAS is compliant with the principle of fairness
- It does not implement racial discrimination

# The debate: Is COMPAS fair?

➤ Is it accurate?



➤ Is it fair to individuals?



➤ Is it fair to groups?

# The case of SAPMOC

- 2000 defendants
  - 1000 blues
  - 1000 greens
- A single predictor:
  - If previous offences then probably recidivate
- Assumption 1
  - previous offenders: 75% recidivate
  - fist-time offenders: 25% recidivate
- Assumption 2
  - Blue: 75% previous offenders
  - Green 25% previous offenders



# SAPMOC Assumptions

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

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

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

# SAPMOC Accuracy

Accuracy	
$(TP+TN)/(TP+FP+TN+FN)$	
Blue	75,0%
Green	75,0%

# SAPMOC FAIRNESS

- Statistical Parity
- Equality of Opportunity
- Calibration
- Conditional Use Error
- Treatment Equality

## Statistical parity



- Each group should have an equal proportion of positives and negatives predictions

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

## Equality of opportunity



- The members of each group, which share the same features, should be treated equally in equal proportion.

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



## Calibration



- The proportion of correct predictions should be equal within each group and with regard to each class.

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

## Conditional use error



- The proportion between FP (FN) and the total amount of positive (negatives) predictions should be equal for the 2 groups.

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

## Treatment equality



- The ratio between errors in positive and negative predictions should be equal in all groups. .

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

# What about SAPMOC/COMPAS?

- 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)
- Individuals fairness vs group fairness

# Consideration on the Fairness in automated decision making

## ➤ 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

# Looking to the 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





Fabio Fossa  
[fabio.fossa@polimi.it](mailto:fabio.fossa@polimi.it)

**MAI4CAREU** Master programmes in Artificial Intelligence 4 Careers in Europe

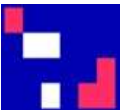
 POLITECNICO DI MILANO



# DRIVING AUTOMATION: AN ETHICAL PERSPECTIVE

 Co-financed by the European Union  
Connecting Europe Facility

This Master is run under the context of Action  
No 2020-EU-IA-0087, co-financed by the EU CEF Telecom  
under GA nr. INEA/CEF/ICT/A2020/2267423



POLITECNICO DI MILANO

# Aims

1. Introduce driving automation and its ethical significance
2. Analyse three ethical issues:
  - Safety
  - Sustainability
  - Responsibility
3. Discuss unavoidable collision moral dilemmas with you



# Driving Automation



# Driving Automation



## SAE J3016™ LEVELS OF DRIVING AUTOMATION

	SAE LEVEL 0	SAE LEVEL 1	SAE LEVEL 2	SAE LEVEL 3	SAE LEVEL 4	SAE LEVEL 5
What does the human in the driver's seat have to do?	You <u>are</u> driving whenever these driver support features are engaged – even if your feet are off the pedals and you are not steering			You <u>are not</u> 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
	<u>Operational Design Domain</u>					
Example Features	<ul style="list-style-type: none"> <li>• automatic emergency braking</li> <li>• blind spot warning</li> <li>• lane departure warning</li> </ul>	<ul style="list-style-type: none"> <li>• lane centering OR</li> <li>• adaptive cruise control</li> </ul>	<ul style="list-style-type: none"> <li>• lane centering AND</li> <li>• adaptive cruise control at the same time</li> </ul>	<ul style="list-style-type: none"> <li>• traffic jam chauffeur</li> </ul>	<ul style="list-style-type: none"> <li>• local driverless taxi</li> <li>• pedals/steering wheel may or may not be installed</li> </ul>	<ul style="list-style-type: none"> <li>• same as level 4, but feature can drive everywhere in all conditions</li> </ul>

# Ethics of Driving Automation

What?

Technical issues:  
Level 5 Autonomy



# Ethics of Driving Automation

Safety



Responsibility Allocation

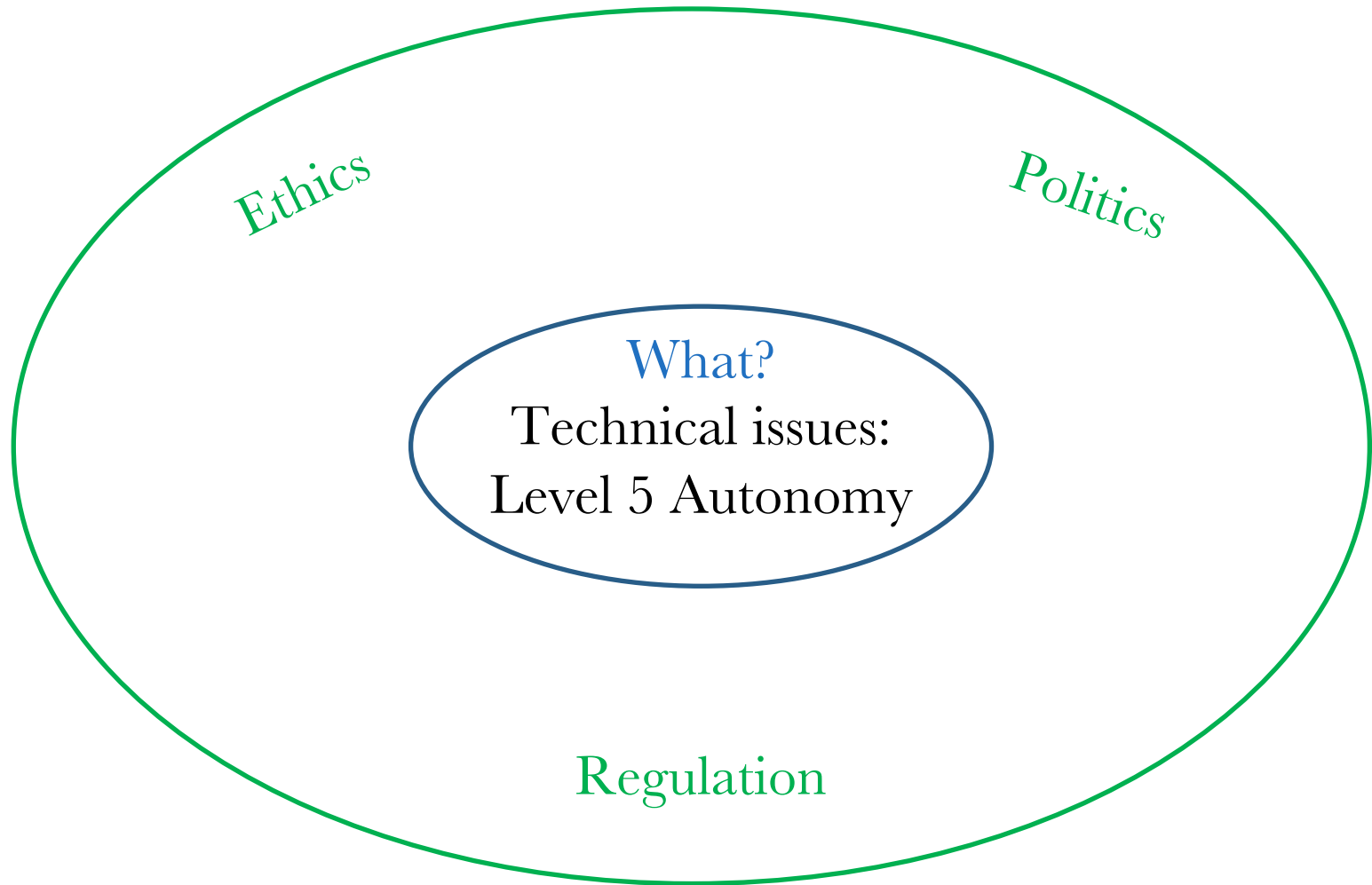


Sustainability

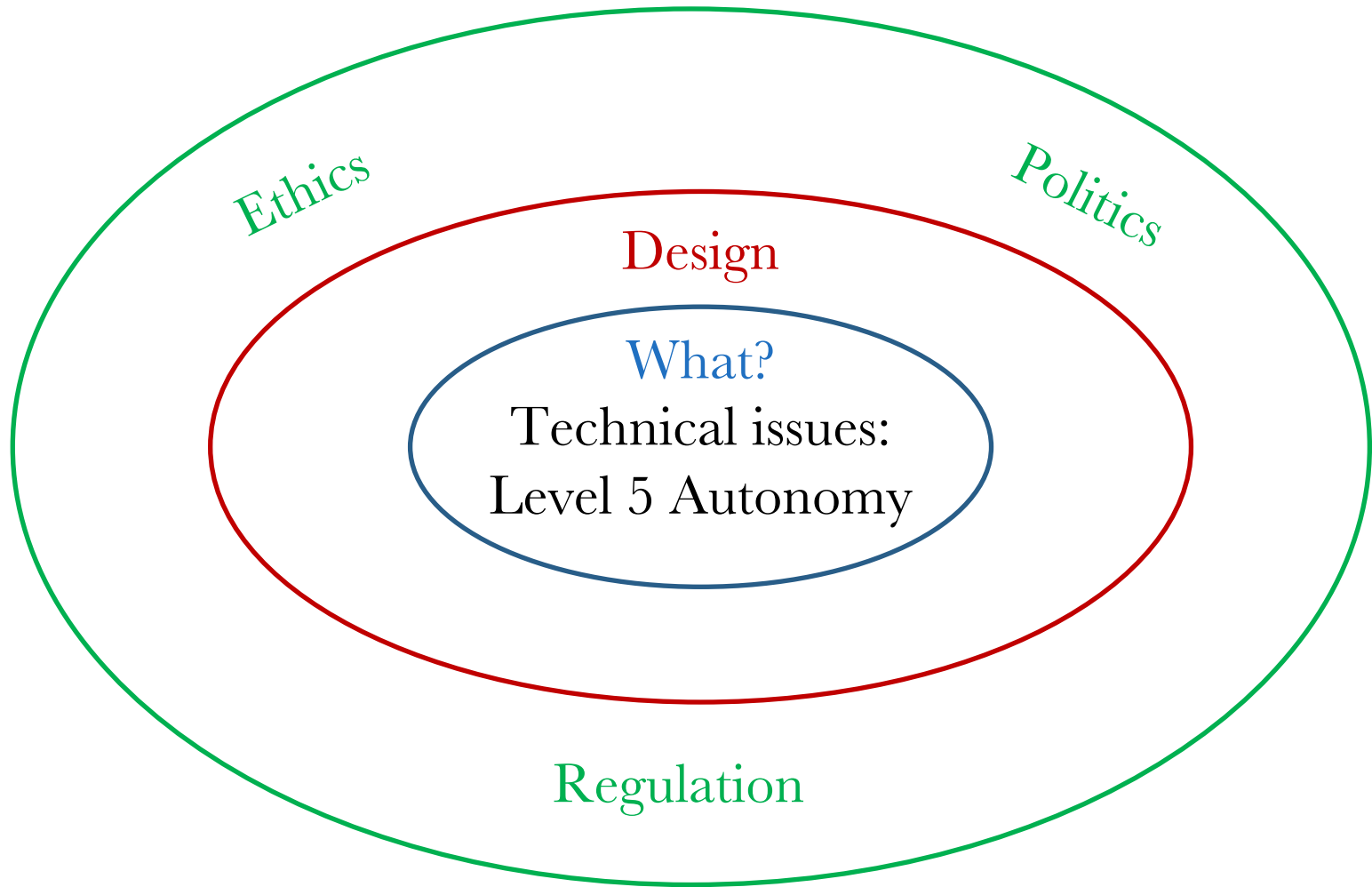




# Ethics of Driving Automation



# Ethics of Driving Automation



↑ **Social Trust** ↑

# Safety



# The Safety Argument



- Traditional problems: ‘usual’ mobility risks
- Artificial Intelligence: new **opportunities!**
- HUGE opportunities – theoretically, at least:
  - ☞ Up to 90% of traffic accidents are caused by human error (text and drive, drunk driving, falling asleep at the wheel, fatigue, road rage, stress...)
  - ☞ 1.3 million deaths per year worldwide



Many collisions will be **avoidable!**

By taking control out of human hands and delivering it to reliable systems, driving automation could dramatically reduce accidents and traffic deaths

# The Safety Argument



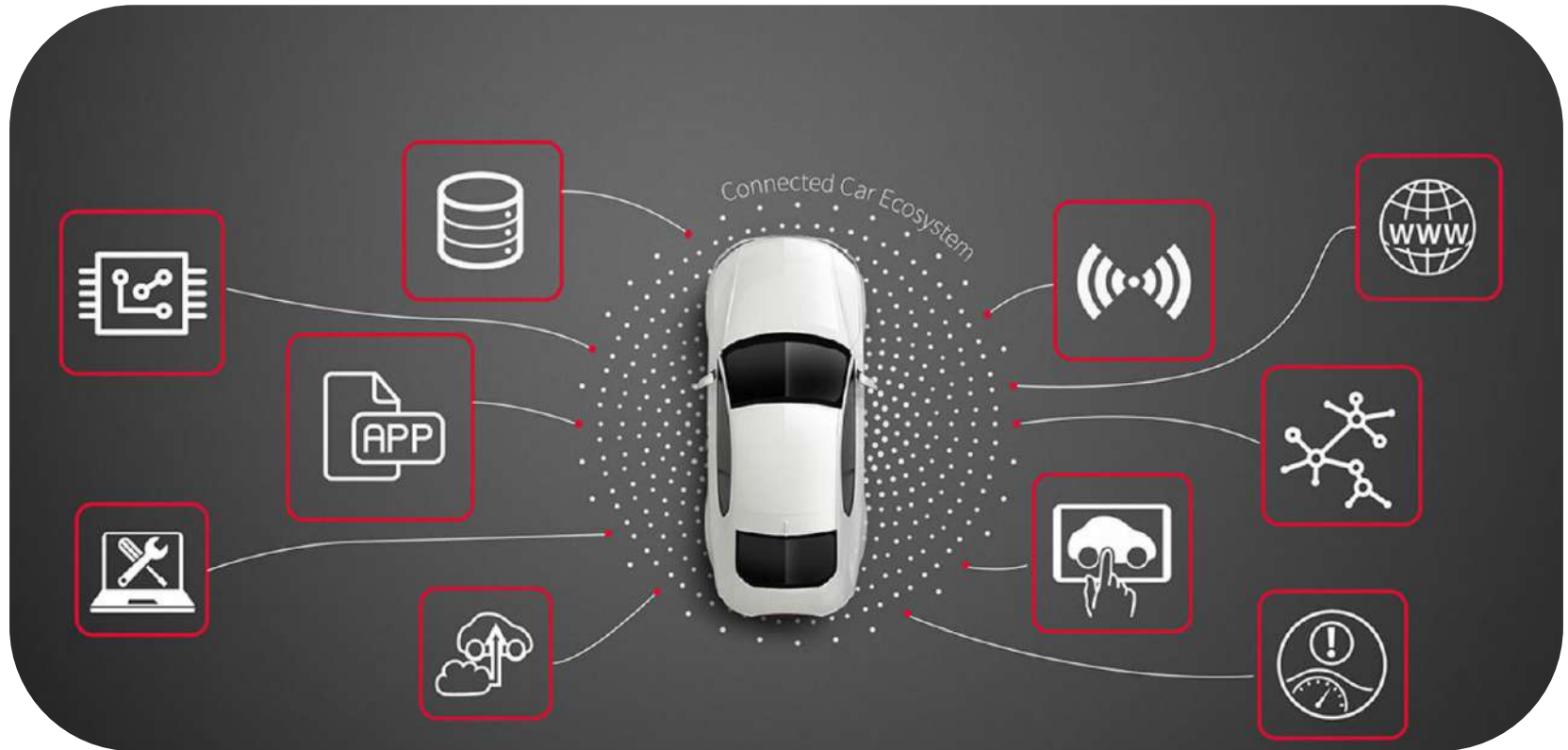
- Traditional problems: ‘usual’ mobility risks
- Artificial Intelligence: new opportunities, **new risks**
- HUGE opportunities – theoretically, at least:
  - ☞ Up to 90% of traffic accidents are caused by human error (text and drive, drunk driving, falling asleep at the wheel, fatigue, road rage, stress...)
  - ☞ 1.3 million deaths per year worldwide



Many collisions will be avoidable!

By taking control out of human hands and delivering it to **reliable** systems, driving automation could dramatically reduce accidents and traffic deaths

# Security & Privacy





# Security



## ***Bogus Satellite Nav Signals Send Autonomous Cars Off the Road***

At the Black Hat security conference, a researcher demonstrated how making tweaks to navigation signals could send a self-driving car careening off the road.



## **Hacking street signs with stickers could confuse self-driving cars**

Subtle or camouflaged optical hacks can change a stop sign into something else.

## **Researcher Hacks Self-driving Car Sensors**

\$60 lidar spoofing device generates fake cars, pedestrians and walls



---

## **Researchers Fool Autonomous Vehicle Systems with Phantom Images**

# Privacy



## How Self-Driving Cars Will Threaten Privacy

Automated vehicles will learn everything about you—and influence your behavior in ways you might not even realize.



### **Self-driving cars: bigger road safety, less privacy**

They are supposed to reduce road casualties, but what else do they entail?

**Self-Driving Cars: Balancing Safety and Data Privacy Considerations**

# Security & Privacy



- Autonomous vehicles pose risks proper of both usual vehicles *and* information systems

# Security & Privacy



- Autonomous vehicles pose risks proper of both usual vehicles *and* information systems
  - ☞ double safety challenge:
    - Make vehicles *safe*: usual stuff (crashworthiness, reliability) but also bugs, software issues, sensor failures, communication problems...
    - Make vehicles *secure*: digital infrastructure liabilities, hijacking, external attacks, data thefts, data leaks...

# Security & Privacy



- Autonomous vehicles pose risks proper of both usual vehicles *and* information systems
  - ☞ double safety challenge:
    - Make vehicles *safe*: usual stuff (crashworthiness, reliability) but also bugs, software issues, sensor failures, communication problems...
    - Make vehicles *secure*: digital infrastructure liabilities, hijacking, external attacks, data thefts, data leaks...
  - ☞ privacy issues:
    - For autonomous vehicles to function properly, a huge quantity of data must be collected, shared, and stored: **personal data** as well!
    - Definitions of privacy and sensible data as involved in AD
    - Privacy protection throughout the entire infrastructure
    - Informed consent (...)

# Safety



## Plus:

- ☞ **Reflect critically** on ethically relevant opportunities and risks
- ☞ Integrate ethical considerations to **design** processes
- ☞ Provide effective **regulation, policy measures,** and **institutional support** to safe mobility



# Sustainability



# Sustainability Narratives



# Sustainability Narratives



## Traffic optimization and new mobility paradigms

- Reduced fuel consumption (air pollution)
- Reduced CO<sub>2</sub> emissions (global warming)
- Reduced land use



# Sustainability Narratives



## Traffic optimization and new mobility paradigms

- Reduced fuel consumption (air pollution)
- Reduced CO<sub>2</sub> emissions (global warming)
- Reduced land use



More wealth

- New business opportunities
- New job opportunities

# Sustainability Narratives



## Traffic optimization and new mobility paradigms

- Reduced fuel consumption (air pollution)
- Reduced CO<sub>2</sub> emissions (global warming)
- Reduced land use

## Increased well-being

- Less time wasted in traffic
- No time wasted in driving
- Improved mobility options
- Inclusivity



## More wealth

- New business opportunities
- New job opportunities

# Sustainability Narratives



**Rebound effects**

Traffic optimization and new mobility paradigms

**Beyond private property?**

- Reduced fuel consumption (air pollution)
- Reduced CO2 emissions (global warming)
- Reduced land use

**Empty trips**

**Data Centres**

Increased well-being

- Less time wasted in traffic
- No time wasted in driving
- Improved mobility options
- Inclusivity



More wealth

- New business opportunities
- New job opportunities



# Sustainability Narratives



**Rebound effects**

Traffic optimization and new mobility paradigms

**Beyond private property?**

- Reduced fuel consumption (air pollution)
- Reduced CO2 emissions (global warming)
- Reduced land use

**Empty trips**

**Data Centres**

**Increased well-being**

- Less time wasted in traffic
- No time wasted in driving
- Improved mobility options
- Inclusivity



More wealth

- New business opportunities
- New job opportunities

**Inequality**

**Technological Unemployment**

# Sustainability Narratives



**Rebound effects**

Traffic optimization and new mobility paradigms

**Beyond private property?**

- Reduced fuel consumption (air pollution)
- Reduced CO2 emissions (global warming)
- Reduced land use

**Empty trips**

**Data Centres**

Increased well-being

- Less time wasted in traffic
- No time wasted in driving
- Improved mobility options
- Inclusivity

**More time to do what?**



**For whom?**

**Inequality**

**Sure?**

More wealth

- New business opportunities
- New job opportunities

**Technological Unemployment**

# Beware of Ethical Innovation Narratives!



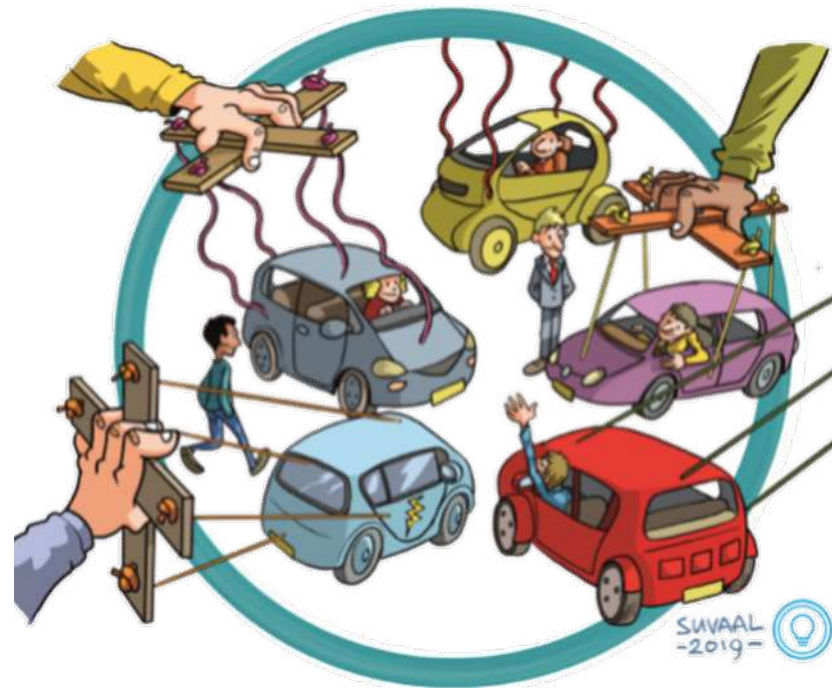
?



Ethical accomplishments depend on how society  
**shapes** driving automation

Moral **commitment** and **responsibility** are crucial

# Responsibility Allocation



# Responsibility Allocation



- Who is to be held responsible for harm caused by accidents where autonomous vehicles are involved?

# Responsibility Allocation



- Who is to be held responsible for harm caused by accidents where autonomous vehicles are involved?
  - The systems themselves?





# Responsibility Allocation



- Who is to be held responsible for harm caused by accidents where autonomous vehicles are involved?
  - The systems themselves? **NO**



# Responsibility Allocation



- Who is to be held responsible for harm caused by accidents where autonomous vehicles are involved?
  - The systems themselves? **NO**
  - passengers?
  - owners?
  - designers/developers?
  - producers?
  - regulators
  - nobody, just insurance system?



# Responsibility Allocation



- Who is to be held responsible for harm caused by accidents where autonomous vehicles are involved?
  - The systems themselves? **NO**
  - passengers?
  - owners?
  - designers/developers?
  - producers?
  - regulators
  - nobody, just insurance system?



# Responsibility Allocation



Tech

## Uber's self-driving operator charged over fatal crash

16 Sept. 2020

## Why Wasn't Uber Charged in a Fatal Self-Driving Car Crash?

17 Sept. 2020

Authorities charged the vehicle's "safety driver" with criminal negligence, but not the company that developed the technology.

## Self-Driving Car Users Shouldn't Be Held Responsible For Crashes, U.K. Report Says

25 Jan. 2022

28 Mar. 2022

Mercedes will accept responsibility for autonomous technology crashes

# Unavoidable Collisions



# Unavoidable Collisions



- Unavoidable collisions have been a primary worry in the ethical debate on autonomous driving
- Q: how should the system handle morally laden situations – i.e., situations where harm is unavoidable *but* can be distributed in different ways? ➔ [Accident-algorithms](#)

rights

duties

non-discrimination

consequences

protect children



sacrifice passengers

sacrifice bystanders

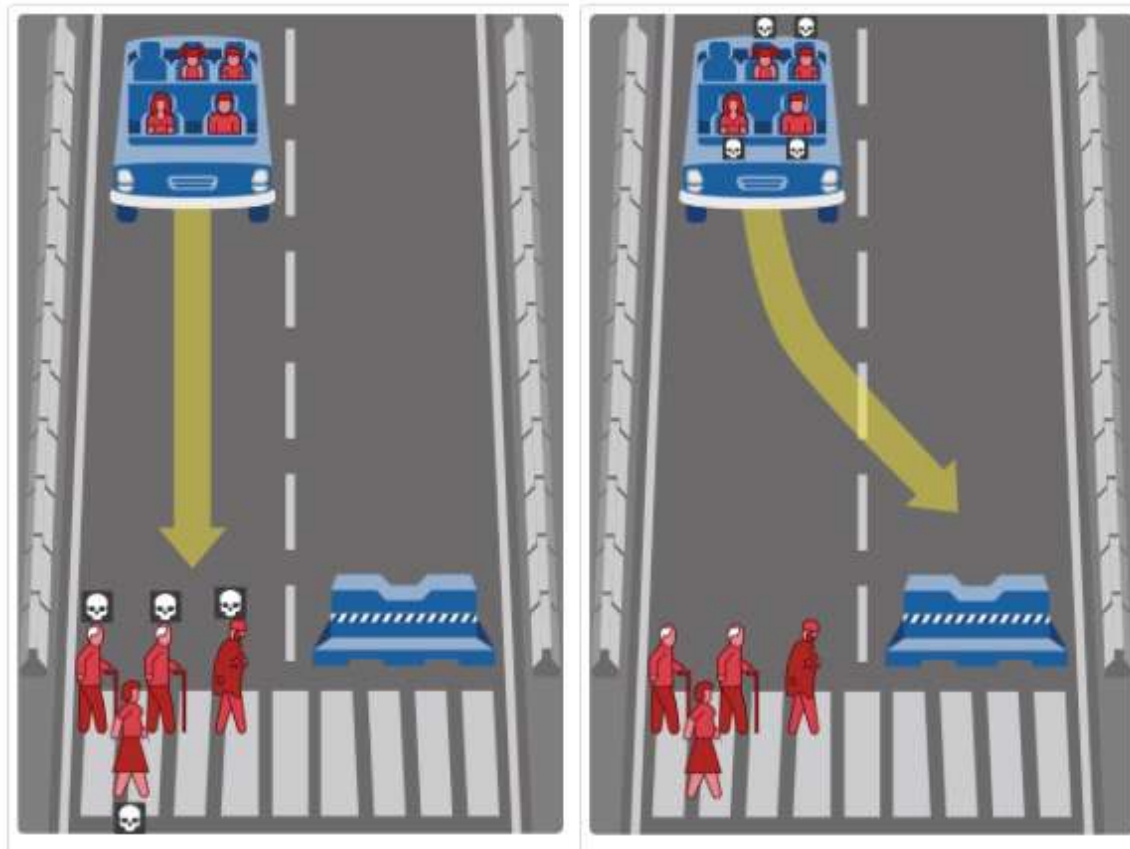
save more lives possible



# Unavoidable Collision



A self-driving car with sudden brake failure is approaching a crosswalk.  
What should it do?



# Unavoidable Collision



A self-driving car with sudden brake failure is approaching a crosswalk.  
What should it do?

## Choice A

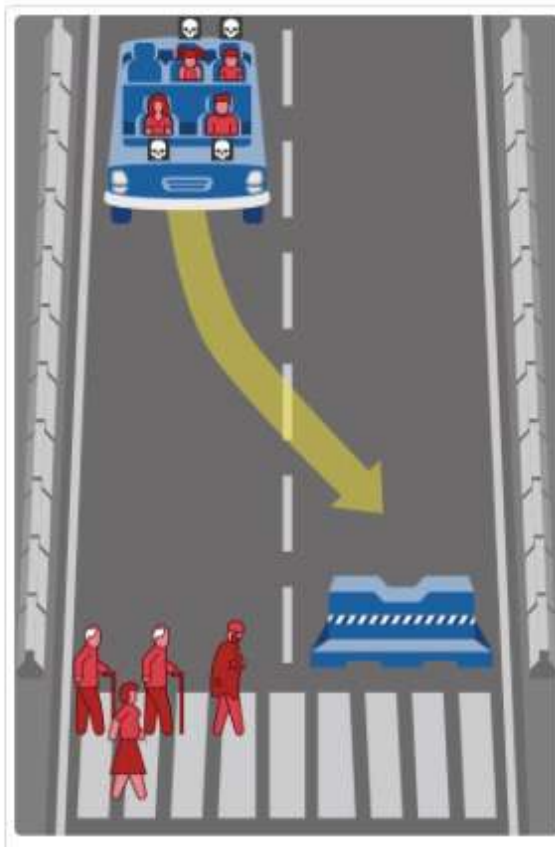
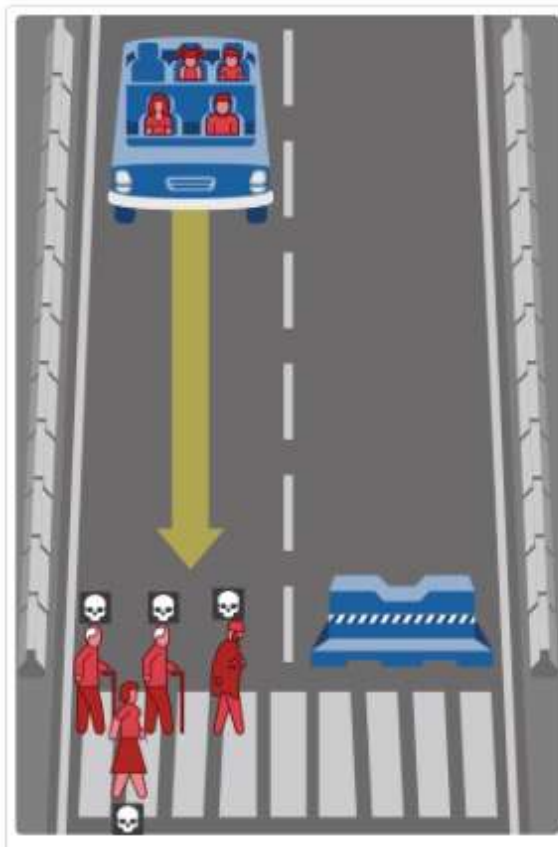
Action:  
don't swerve

Consequences:  
Pedestrians die

2 elderly men  
1 homeless person  
1 fat woman

Passengers safe:

1 man  
1 woman  
1 little girl  
1 little boy



# Unavoidable Collision



A self-driving car with sudden brake failure is approaching a crosswalk.  
What should it do?

## Choice A

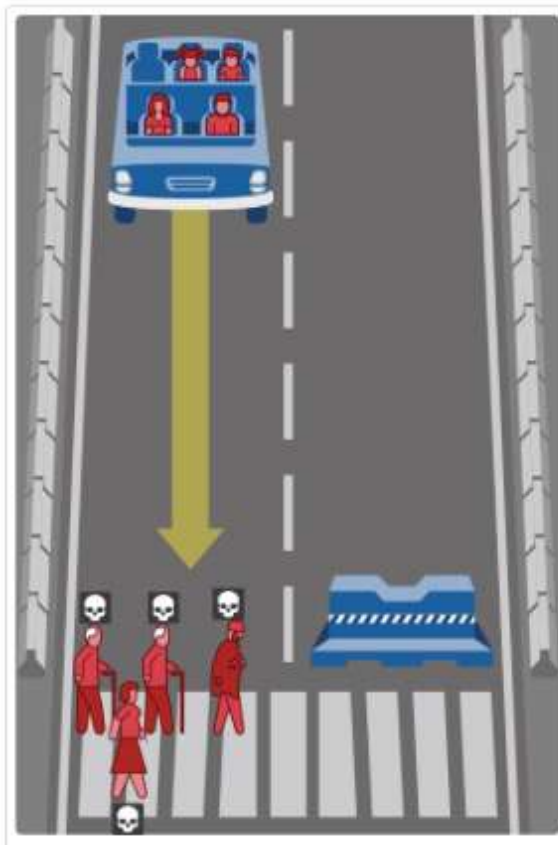
Action:  
don't swerve

Consequences:  
Pedestrians die

2 elderly men  
1 homeless person  
1 fat woman

Passengers safe:

1 man  
1 woman  
1 little girl  
1 little boy



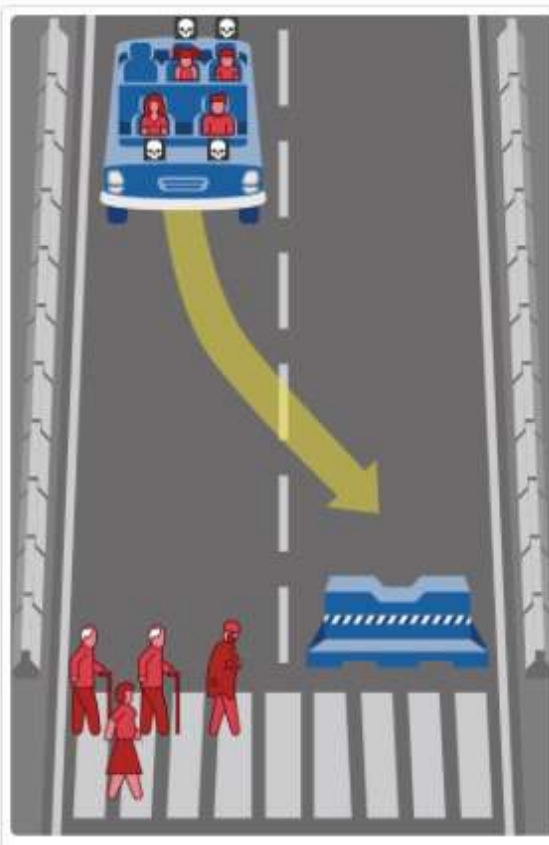
## Choice B

Action:  
swerve

Consequences:  
Passengers die

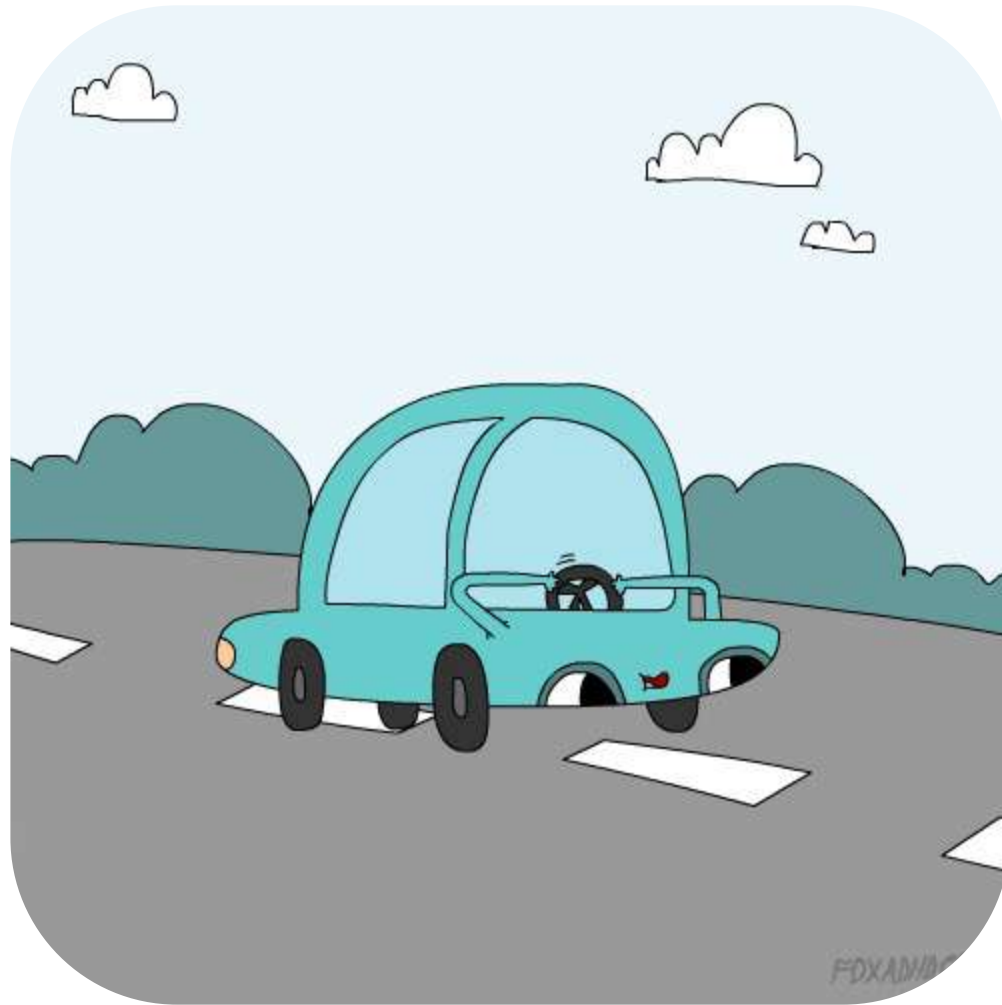
1 man  
1 woman  
1 little girl  
1 little boy

Pedestrians safe:  
2 elderly men  
1 homeless person  
1 fat woman



Thank you for  
your attention!

Questions?






```
for object to mirror_mod.mirror_object
operation == "MIRROR_X":
mirror_mod.use_x = True
mirror_mod.use_y = False
mirror_mod.use_z = False
operation == "MIRROR_Y":
mirror_mod.use_x = False
mirror_mod.use_y = True
mirror_mod.use_z = False
operation == "MIRROR_Z":
mirror_mod.use_x = False
mirror_mod.use_y = False
mirror_mod.use_z = True
```

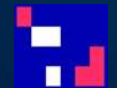
# Ethics of Filtering

Andrea Loreggia

**MAI4CAREU** | Master programmes in Artificial Intelligence 4 Careers in Europe

 Co-financed by the European Union  
Connecting Europe Facility

This Master is run under the context of Action No 2020-EU-IA-0087, co-financed by the EU CEF Telecom under GA nr. INEA/CEF/ICT/A2020/2267423



# Introduction

- 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.



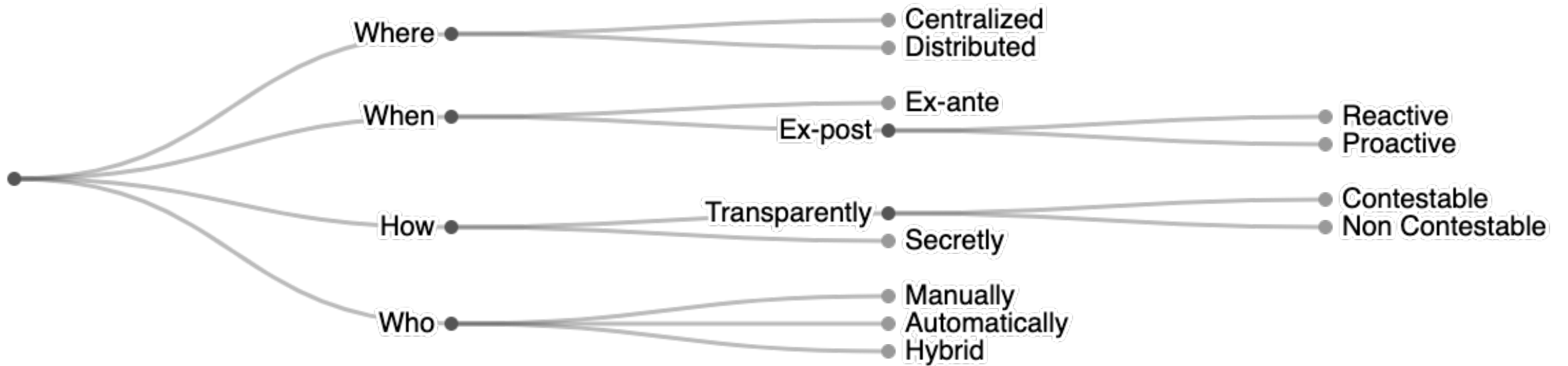
# What is moderation?

- Moderation is the active governance of platforms meant to ensure interactions among the users that are:
  - Productive
  - Pro-social
  - Lawful

# Why filtering?

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

# Taxonomy



# Taxonomy - Where

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

# Taxonomy - When

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

# Taxonomy - When

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



# Taxonomy - How

- *Transparent filtering*, which provides information on the exclusion of items from the platform.
- *Secret filtering*, which does not provide any information about the operation.

# Taxonomy - How

- *Transparent filtering*, which provides information on the exclusion of items from the platform.
  - *Contestable filtering*. The platform provides uploaders with ways to contest the outcome of the filtering, and to obtain a new decision on the matter.
  - *Non-contestable filtering*. No remedy is available to the uploaders.
- *Secret filtering*, which does not provide any information about the operation.

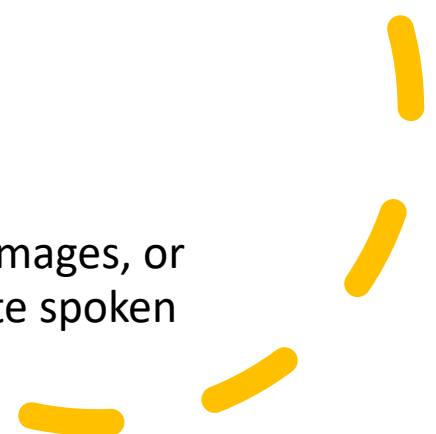
# Taxonomy - Who

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

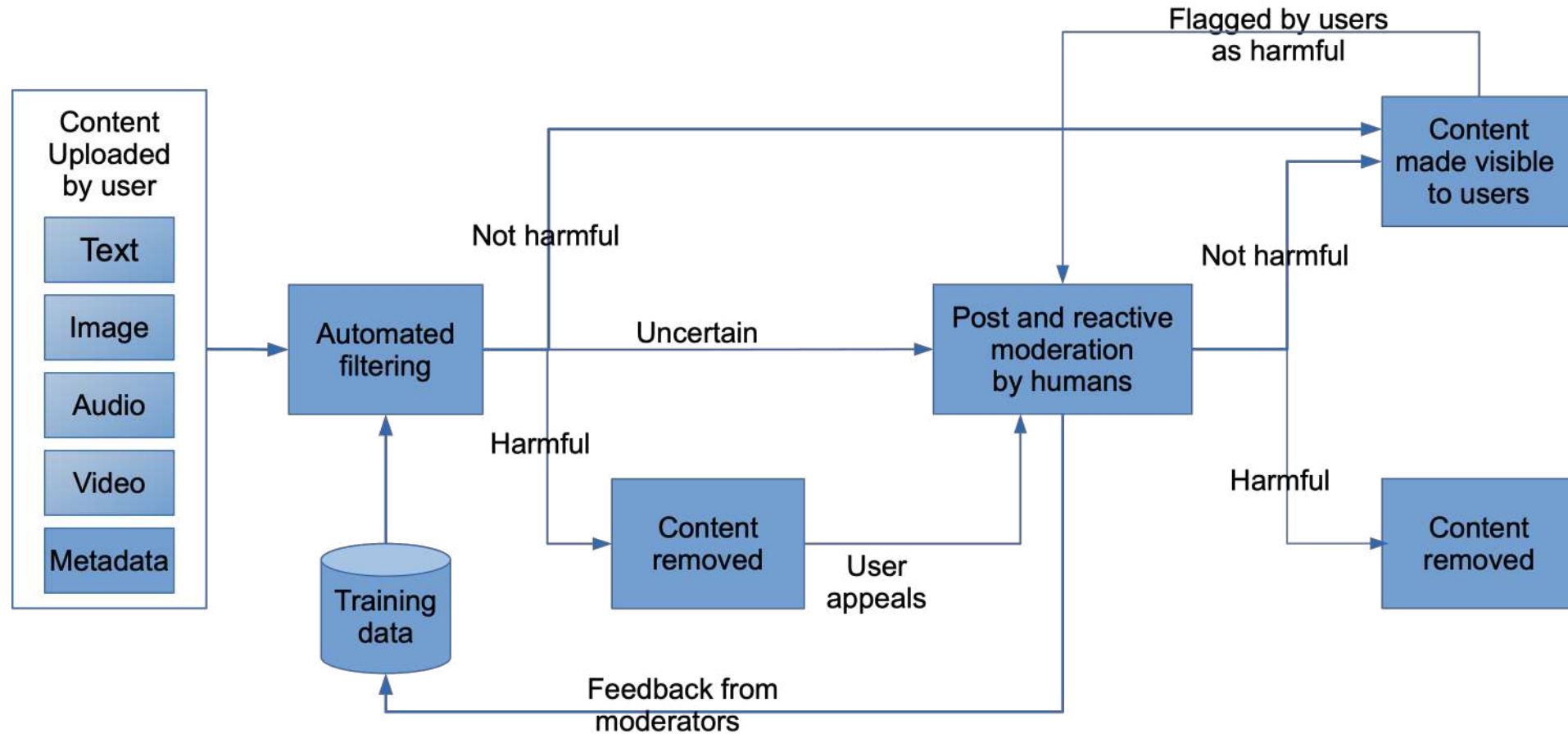


# Different Media

- Metadata searching, hashing, and fingerprinting → to identify copies of known digital works;
- Blacklisting → to find unwanted expressions;
- NLP → to address meaning and context;
- Multiple AI techniques → to identify unwanted images, or combinations of text and images, and to translate spoken language into text.




# How it works



# Epic Fails

INDEPENDENT

Subscribe LOGIN



The Little Mermaid statue is one of Denmark's best-loved sights (ODD ANDERSEN/AFP/Getty Images)

**FACEBOOK REMOVES IMAGE OF COPENHAGEN'S LITTLE MERMAID STATUE FOR BREAKING NUDITY RULES**

CNN travel VIDEO Q



Curiosità e scorci di Bologna

Enza Barbara FANPAGE

Facebook banned Neptune statue photo for being 'explicitly sexual'

Sara Delgrossi and Lauren Said-Moorhouse, CNN • Updated 5th January 2017



# Epic Fails

ISSIE LAPOWSKY BUSINESS 03.15.2019 01:50 PM

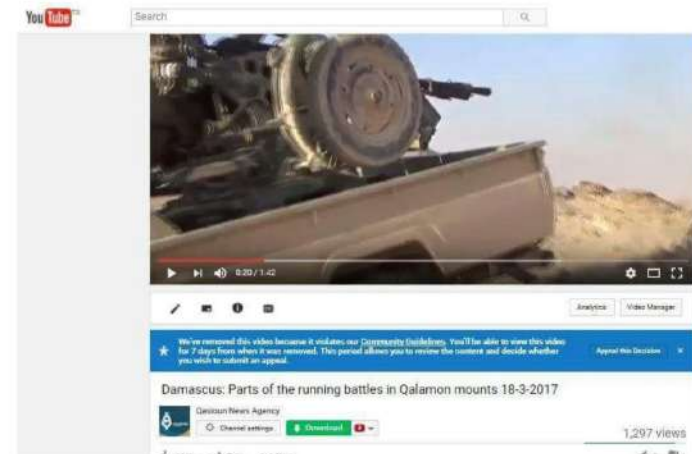
## Why Tech Didn't Stop the New Zealand Attack From Going Viral

Video from mosque shootings in Christchurch popped up on Facebook, Reddit, Twitter, and YouTube, showing the limits of social media moderation.



Creazione di una connessione protetta in...

## YouTube Removes Videos Showing Atrocities in Syria



A takedown notice issued by YouTube on a video of the Syrian conflict. YouTube

# Santa Clara Principles

---



**Companies should publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines.**

# Santa Clara Principles

---



**Companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension.**

# Santa Clara Principles

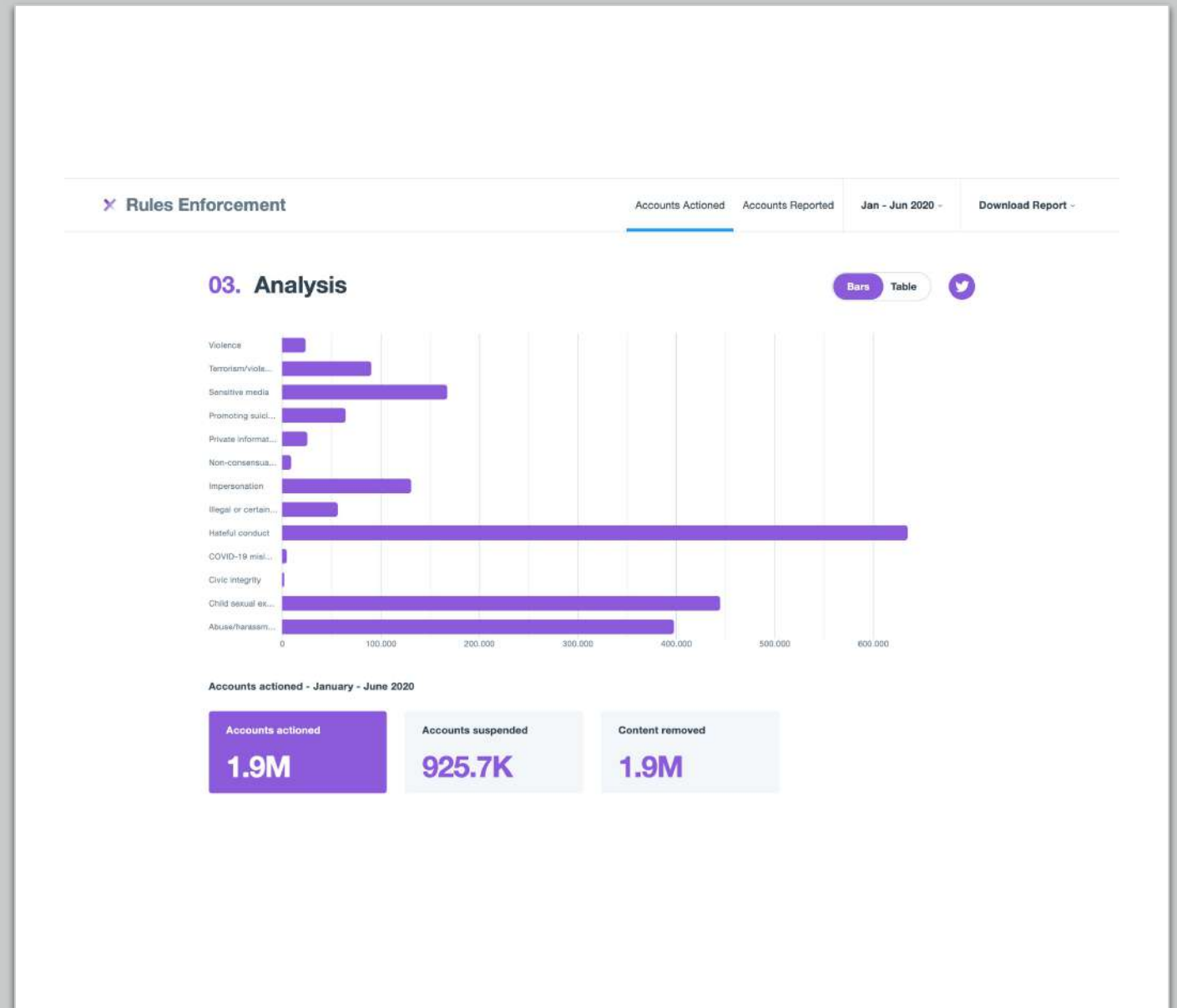
---



**Companies should provide a meaningful opportunity for timely appeal of any content removal or account suspension.**

# Transparency

- Example from Twitter transparency report







Issues on

- Filter bubbles
- Echo chambers
- Censorship
- Fake news



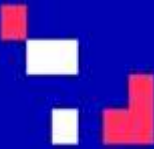




**UNIMORE** ALMA MATER STUDIORUM  
UNIVERSITÀ DEGLI STUDI DI  
MODENA E REGGIO EMILIA



claudette.eui.eu



# Text analytics in the legal domain:

the case of contracts and privacy policies in Claudette

**Francesca Lagioia, PhD**







[francesca.lagioia@eui.eu](mailto:francesca.lagioia@eui.eu), [francesca.lagioia@unibo.it](mailto:francesca.lagioia@unibo.it)



# The Claudette Team

Law Dept@eui

Cirsfid@unibo

 Hans-Wolfgang Micklits	 Kasper Drazewski	 Giacomo Tagiuri	 Agnieszka Jablonowska	 Francesca Lagioia	 Giovanni Sartor
---	---	---	--	--	--



Przemyslaw Palka

Law School@yale



Marco Coppi

DISMI@unimore

 Paolo Torroni	 Federico Ruggeri	 Andrea Galassi
---	--	--

DISI@unibo



Ruta Liepina

Law Dept@Maastricht



claudette.eui.eu



# How to empower consumers?

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

*“An opposing exercise of power is the principal solvent of economic power, the basic defense against its exercise in economic affairs”. Ken Galbraith*

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

# CLAUDETTE.eui.eu

Automatically detect **potentially** unfair clauses in Terms of Services and Privacy Policies

- Consumers agree but don't read
- NGOs have competence to control but lack resources
- Business keeps using unlawful clauses



# Terms of Service (ToS): The Training Set

## The ToS Corpus

### WHERE DID WE START?

... 50 ToS (manually annotated)...

7,090 sentences, 787 of which (11.1%) were labeled as positive, thus containing a potentially unfair clause.

### WHERE ARE WE NOW?

... 100 ToS (manually annotated)...





# Part 1: Unfair Contract Terms Law and Practice

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.

Bottom-line: there are some types of clauses that traders are prohibited from using in the contracts.

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

Type of clause	Symbol	# 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	<ltl>	296	49
Unilateral termination	<ter>	236	48
Consent by using	<use>	117	48
Privacy included	<pinc>		

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

[claudette.eui.eu](http://claudette.eui.eu)

# Consent by using Clause

If a clause states that the consumer is bound by the terms of service simply by visiting the website or by downloading the app, or by using the service: **potentially unfair**

## A **potentially unfair** consent by using clause (Airbnb):

`<use2>By accessing or using the Airbnb Platform, you agree to comply with and be bound by these Terms of Service.</use2>`

## 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>`

# Jurisdiction Clause

## Where a dispute will be adjudicated?

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

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

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>
```

# Limitation of Liability

For what actions/events the provider claims they will not be liable?

If stating that the provider may be liable: **clearly fair**

If stating that the provider will never be liable for any action taken by other people// damages incurred by the computer because of malware // When contains a blanket phrase like “to the fullest extent permissible by law”: **potentially unfair**

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

# Limitation of Liability

For what actions/events the provider claims they will not be liable?

A **fair liability** clause (World of Warcraft):

```
<ltd1>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.</ltd1>
```

A **potentially unfair** limitation of liability clause (9gag):

```
<ltd2>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.</ ltd2>
```





# Limitation of Liability

For what actions/events the provider claims they will not be liable?

A **potentially unfair** limitation of liability clause (Truecaller):

<1td2>To the maximum extent permitted by applicable law, you expressly agree that truecaller shall in no event be liable for any direct, indirect, special, incidental, consequential or exemplary damages, including but not limited to damages for loss of profits, data and goodwill, arising out of the use or inability to use the services or the content, even if advised of the possibility of such damages</1td2>

# Limitation of Liability

For what actions/events the provider claims they will not be liable?

A **clearly unfair** limitation of liability clause (Rovio):

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, [...] and, to the extent permitted by law, damages for **personal injuries**, [...] whether or not Rovio, Rovio's licensors or channel partners have been advised of the possibility of such damages.

# Privacy Included

Whenever a clause states (or it might be possible to assume) that the consumer consents to the privacy policy simply by using the service: **potentially unfair**

A potentially unfair clause (**DeviantArt**):

<use2>By using our Service, you agree to be bound by Section I of these Terms ("General Terms"), which contains provisions applicable to all users of our Service, including visitors to the DeviantArt website (the "Site").</use2> [...]

<pinc2>The terms of DeviantArt's privacy policy are incorporated into, and form a part of, these Terms.</pinc2>

# Arbitration Clause

Is arbitration mandatory before the case can go to court?

If arbitration is fully optional: **clearly fair**

If arbitration should take place in a state other than the state of consumer's residence and/or be based on arbiter's discretion (i.e. not on law): **clearly unfair**

All other arbitration clauses: **potentially unfair**

# Arbitration Clause

Is arbitration mandatory before the case can go to court?

A **clearly unfair** arbitration clause (Rovio):

<a3>Any dispute, controversy or claim arising out of or relating to this EULA or the breach, termination or validity thereof shall be finally settled at Rovio's discretion (i) at your domicile's competent courts; or (ii) by arbitration in accordance with the Rules for Expedited Arbitration of the Arbitration Institute of the Finland Chamber of Commerce. The arbitration shall be conducted in Helsinki, Finland, in the English language.</a3>

# An example from the Instagram Terms of Service

We reserve the right, in our sole discretion, to change these Terms of Use ("Updated Terms") from time to time.

Unless we make a change for legal or administrative reasons, we will provide reasonable advance notice before the Updated Terms become effective. You agree that we may notify you of the Updated Terms by posting them on the Service, and that your use of the Service after the effective date of the Updated Terms (or engaging in such other conduct as we may reasonably specify) constitutes your agreement to the Updated Terms.





<lt2>TO THE EXTENT PERMITTED BY LAW, THE TOTAL LIABILITY OF GOOGLE, AND ITS SUPPLIERS AND DISTRIBUTORS, FOR ANY CLAIMS UNDER THESE TERMS, INCLUDING FOR ANY IMPLIED WARRANTIES, IS LIMITED TO THE AMOUNT YOU PAID US TO USE THE SERVICES (OR, IF WE CHOOSE, TO SUPPLYING YOU THE SERVICES AGAIN).</lt2>

<lt2>IN ALL CASES, GOOGLE, AND ITS SUPPLIERS AND DISTRIBUTORS, WILL NOT BE LIABLE FOR ANY LOSS OR DAMAGE THAT IS NOT REASONABLY FORESEEABLE.</lt2>

### Business uses of our Services

If you are using our Services on behalf of a business, that business accepts these terms. It will hold harmless and indemnify Google and its affiliates, officers, agents, and employees from any claim, suit or action arising from or related to the use of the Services or violation of these terms, including any liability or expense arising from claims, losses, damages, suits, judgments, litigation costs and attorneys' fees.

### About these Terms

<ch2>We may modify these terms or any additional terms that apply to a Service to, for example, reflect changes to the law or changes to our Services. You should look at the terms regularly.</ch2> We'll post notice of modifications to these terms on this page. We'll post notice of modified additional terms in the applicable Service. Changes will not apply retroactively and will become effective no sooner than fourteen days after they are posted. However, changes addressing new functions for a Service or changes made for legal reasons will be effective immediately. If you do not agree to the modified terms for a Service, you should discontinue your use of that Service.

If there is a conflict between these terms and the additional terms, the additional terms will control for that conflict.

These terms control the relationship between Google and you. They do not create any third party beneficiary rights.

If you do not comply with these terms, and we don't take action right away, this doesn't mean that we are giving up any rights that we may have (such as taking action in the future).

If it turns out that a particular term is not enforceable, this will not affect any other terms.

<law2>The laws of California, U.S.A., excluding California's conflict of laws rules, will apply to any disputes arising out of or relating to these terms or the Services.</law2> <j3>All claims arising out of or relating to these terms or the Services will be litigated exclusively in the federal or state courts of Santa Clara County, California, USA, and you and Google consent to personal jurisdiction in those courts.</j3>

# The Machine Learning Methodology

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**: to what category does the unfair clause belong?

## Approaches

- Bag of Words (BoW): leverages on the lexical information in sentences
- Tree kernels: leverages on grammatical structure of sentences
- Convolutional Neural Networks, SVM, etc.

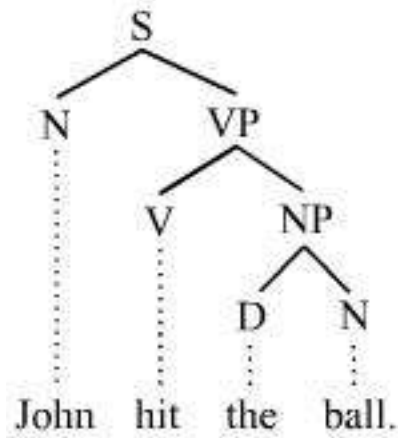
# The Bag of Words Model (BoW)

- Build to leverage the lexical information in sentences
- Each word is a feature
- Each sentence is represented as a vector of features, as large as the dimension of vocabulary in the corpus (also bigrams)
- Each feature is either zero (if the word does not appear in the sentence), or different than zero (if it appears)
- We feed VECTORS to Support Vector Machines (SVMs)

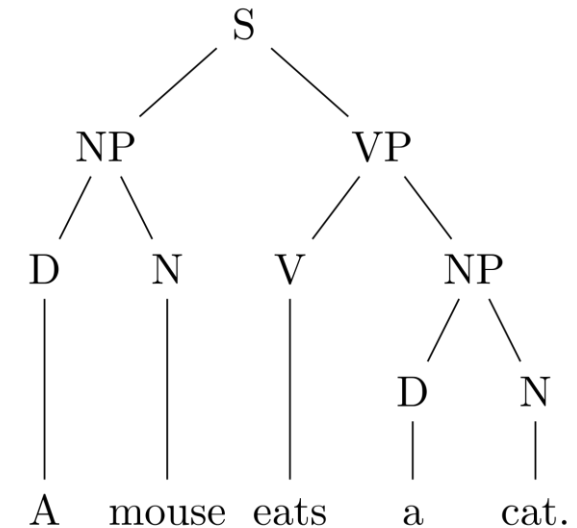
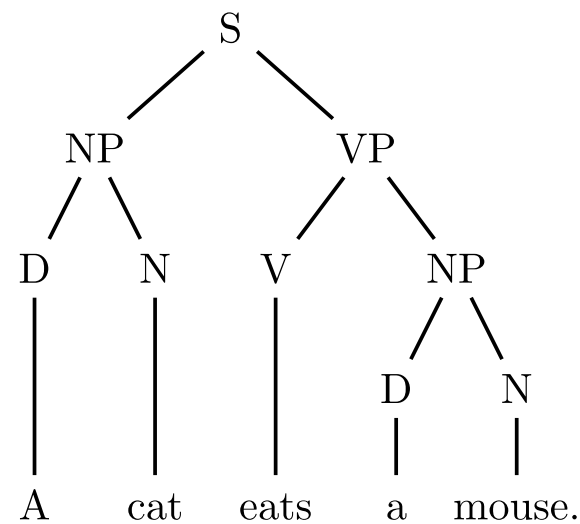
TEXT	VOCABULARY	VECTORS
<i>1. It was the best of times,</i>	It	[it, was, the, best, of, times, ...]
<i>2. it was the worst of times,</i>	was	1. [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]
<i>3. it was the age of wisdom,</i>	the	2. [1, 1, 1, 0, 1, 1, 1, 0, 0, 0]
<i>4. it was the age of foolishness,</i>	Best	3. [1, 1, 1, 0, 1, 0, 0, 1, 1, 0]
	of	4. [1, 1, 1, 0, 1, 0, 0, 1, 0, 1]
	Times	
	<b>Worst</b> <- ONE OCCURRENCE	
	[...]	

# The Tree Kernels Model

- A method for representing the **semantic** structure of sentences (Constituency-based parse tree)
- A method for comparing tree graphs to each other, allowing us to get quantifiable measurements of their similarities or differences
- A TK consists of a similarity measure between two trees, which takes into account the number of common substructures, known as fragments
- More sophisticated, lately proven to be effective in argumentation mining



Constituency-based parse tree



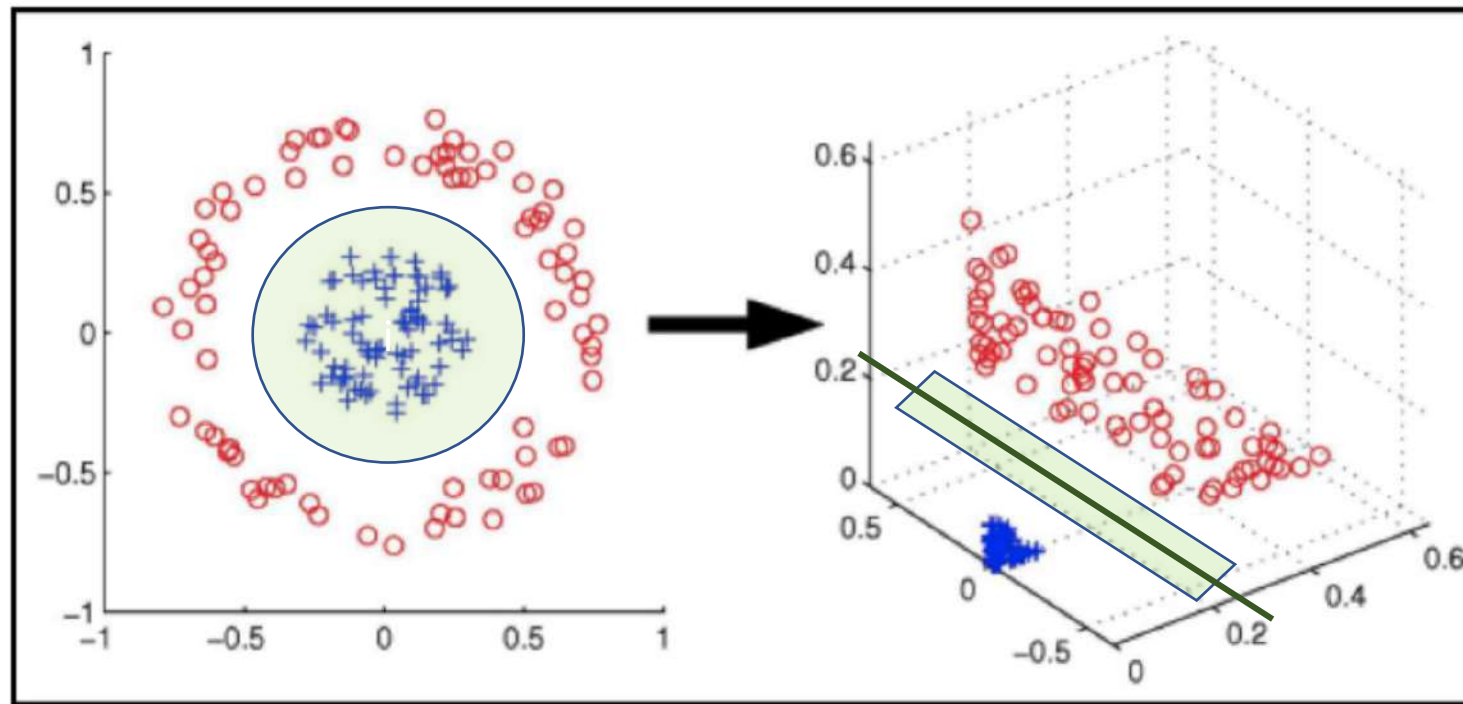


# Supported Vector Machines (SVM)

Once we've defined our mathematical space (set of vectors mapping sentences) SVM can be used to detect whether a new clause is fair or not, and the category it belongs to.

SVM is a **binary classifier**. It builds an **hyper plane classifier** (in an augmented space).

SVMs calculate a **maximum-margin boundary** that leads to a homogeneous partition of all data points. This classifies a SVM as a **maximum margin classifier**.



# Data representation and Ensemble Methods

The problem is formulated as a binary classification task where

- the positive class is either the union of all potentially unfair sentences
- or the set of potentially unfair clauses of a single category

Results of each configuration have been collected and compared to see which one performs better. The **better performance** is an **ensemble**.

C1: SVM exploiting BoW

C2: SVM exploiting TK for sentence representation

C3: SVM for collective classification of sentences in a document (BoW+TK)



Voting: if 2 out of 3 predict positive sentence

The input sentence is classified as potential unfair



# Experiments

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

## 3 Metrics

**Precision:** fraction of positive predictions, actually labelled as positive

**Recall:** fraction of positive examples that are correctly detected

**F1:** harmonic mean between precision and recall

**Baselines for comparison:** random classifier

# Experimental Results

Performance: Training set size = 50 Tos

Method	$P$	$R$	$F_1$
SVM—single model	0.729	<b>0.830</b>	0.769
SVM—combined model	0.798	0.782	0.781
Tree kernels	0.777	0.718	0.739
Convolutional neural networks	0.729	0.739	0.722
Long short-term memory networks	0.696	0.723	0.698
SVM-HMM—single model	0.759	0.778	0.758
SVM-HMM—combined model	<b>0.859</b>	0.687	0.757
Ensemble (C1+C2+C3+C6+C7)	0.826	0.797	<b>0.805</b>
Random baseline	0.125	0.125	0.125
Always positive baseline	0.123	1.000	0.217

Best performing: **Ensemble**

[claudette.eui.eu](http://claudette.eui.eu)

# Experimental Results

Claudette correctly detected around **80%** of the **potentially unfair** clauses in each category, ranging from a **minimum 72.7%** in the case of arbitration clauses, **up to 89.7%**, as in the case of jurisdiction clauses.

Tag	Precision	Recall	F <sub>1</sub>
Arbitration	0.832	0.814	0.823
Unilateral change	0.832	0.814	0.823
Content removal	0.713	0.780	0.745
Jurisdiction	1.000	0.941	0.970
Choice of law	0.984	0.886	0.932
Limitation of liability	0.961	0.905	0.932
Unilateral termination	0.786	0.932	0.853
Contract by using	0.949	0.957	0.953

# An online server

**CLAUDETTE**

An automated detector of potentially unfair clauses

Copy your text here

[Submit](#)

[About](#) [Cite](#) [Contact](#)

# CLAUDETTE

## An Automated Detector of Potentially Unfair Clauses

Claudette found 3 potentially unfair clauses (displayed in **bold**) out of 295 sentences.  
Below you can find a summary of the detected clauses, possibly linked to the most plausible rationales.

### Potentially unfair clause #1

**EXCEPT FOR CERTAIN TYPES OF DISPUTES MENTIONED IN THE ARBITRATION CLAUSE , YOU AND HEADSPACE AGREE THAT DISPUTES RELATING TO THESE TERMS OR YOUR USE OF THE PRODUCTS WILL BERESOLVED BY MANDATORY BINDING ARBITRATION , AND YOU WAIVE ANY RIGHT TO PARTICIPATE IN A CLASS-ACTION LAWSUIT OR CLASS-WIDE ARBITRATION .**

Unfairness categories: **Arbitration**

[Hide/show rationales](#)

### Potentially unfair clause #2

**1.4 CHANGES TO TERMS** Headspace reserves the right to change or update these Terms , or any other of our policies or practices , at any time , and will notify users by posting such changed or updated Terms on this page .

Unfairness categories: **Unilateral Change**

[Hide/show rationales](#)

### Potentially unfair clause #3

**Your continued use of the Products constitutes your agreement to abide by the Terms as changed .**

Unfairness categories: **Contract by Using**

[Hide/show rationales](#)

# CLAUDETTE

## An Automated Detector of Potentially Unfair Clauses

### Potentially unfair clause #1

**EXCEPT FOR CERTAIN TYPES OF DISPUTES MENTIONED IN THE ARBITRATION CLAUSE , YOU AND HEADSPACE AGREE THAT DISPUTES RELATING TO THESE TERMS OR YOUR USE OF THE PRODUCTS WILL BERESOLVED BY MANDATORY BINDING ARBITRATION , AND YOU WAIVE ANY RIGHT TO PARTICIPATE IN A CLASS-ACTION LAWSUIT OR CLASS-WIDE ARBITRATION .**

Unfairness categories: **Arbitration**

[Hide/show rationales](#)

### Potentially unfair clause #2

**1.4 CHANGES TO TERMS** Headspace reserves the right to change or update these Terms , or any other of our policies or practices , at any time , and will notify users by posting such changed or updated Terms on this page .

Unfairness categories: **Unilateral Change**

[Hide/show rationales](#)

The clause is potentially unfair for **Unilateral Change** since the provider has the right for unilateral change of the contract, services, goods, features for any reason at its full discretion, at any time (score = 0.834)

### Potentially unfair clause #3

**Your continued use of the Products constitutes your agreement to abide by the Terms as changed .**

Unfairness categories: **Contract by Using**

[Hide/show rationales](#)





Human Legal experts are able to recognize potentially unfair clauses thanks to their **background knowledge** of the domain.

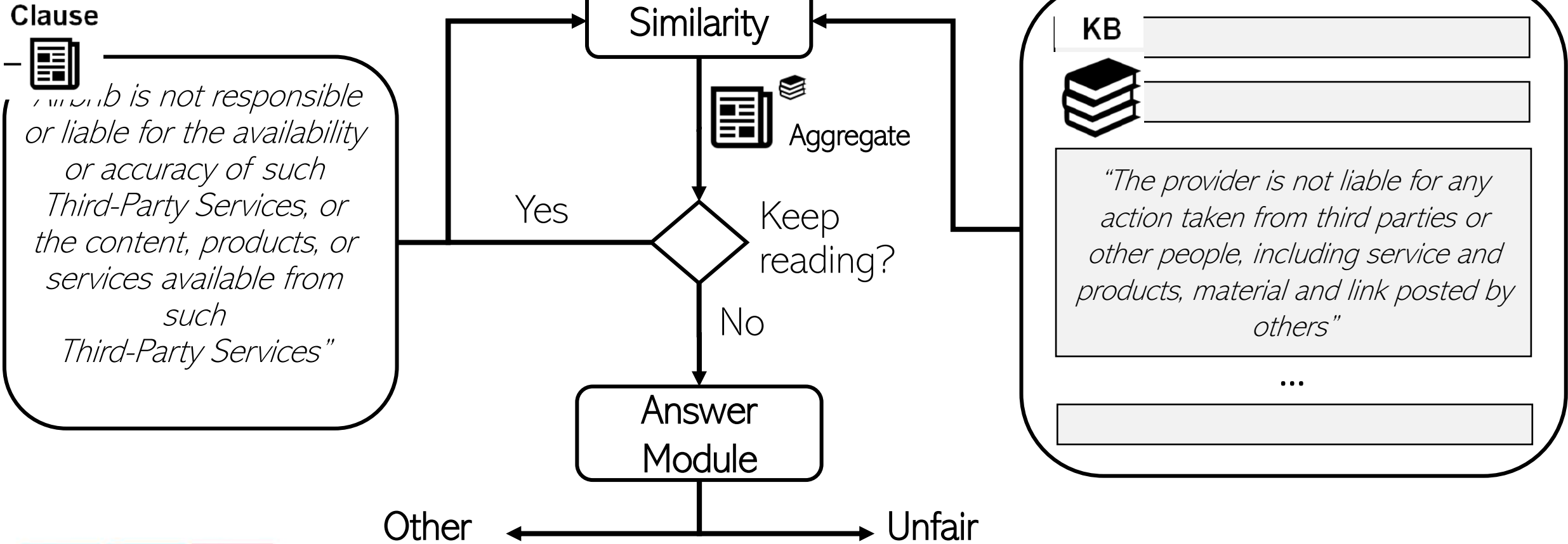
- Rely on intuitions, trained on experience with relevant examples
- Able to explain their intuitions of unfairness, provide reasons why a clause is unfair (**Legal Rationales**), and use rationales to guide such intuitions
- Appealing to their background knowledge (e.g. Standards, Rules and Principles, Judicial precedents) as support for reasoning

# Memory-Augmented Neural Networks

- Process input and **store** the information in some **memory**
- Understand **pieces of knowledge** relevant to a given **query**
- Retrieve **concepts** from memory
- Combine **memory and query** to make a prediction

# Exploiting Knowledge for Unfairness Identification

## Esperimental Setup



# CLAUDETTE meets GDPR

Moving to **privacy policies**: what is different?

Ensure some information is **present** and **complete** (i.e., compliant with articles 13-14 of GDPR)

Detect problematic clauses for **data processing**

Detect **vague** language

# CLAUDETTE meets GDPR

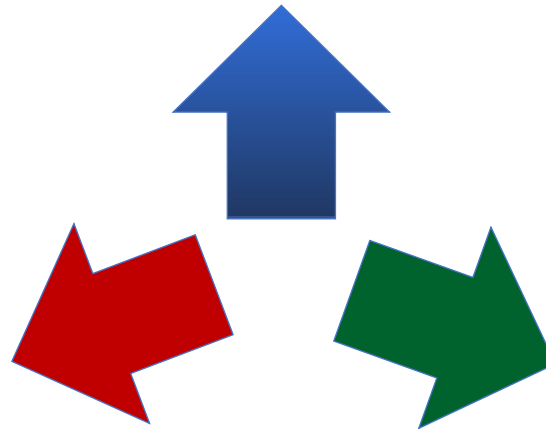
The Golden Standard: Lawfulness Fairness Transparency

## Comprehensiveness of information

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

## Clarity of expression

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



## Substantive compliance

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

Different Levels of Achievement: Optimal and Suboptimal (questionable or insufficient)

# Comprehensiveness of information

23 categories (GDPR art 13 and 14)

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>
...	<...>



# Purposes of the processing for which the personal data are intended

13(1)(c) and 14(1)(c))

Clauses where the purposes of the processing are exhaustive and not vague: **fully informative**

In other cases (e.g. when a clause only provides examples): **insufficiently informative**

WhatsApp Privacy Policy (last updated on 24 April 2018)

`<purp2> WhatsApp must receive or collect some information to operate, provide, improve, understand, customize, support, and market our Services, including when you install, access, or use our Services.</purp2>`



# The Categories of personal data concerned

Clauses where the categories of personal data are comprehensively specified and not vague: **fully informative**

In other cases (e.g. when a clause only provides examples): **insufficiently informative**

Google Privacy Policy (last updated on 25 May 2018)

```
<cat1>We collect information about your location when you use our services, which helps us offer features like driving directions for your weekend getaway or showtimes for movies playing near you.</cat1>
```



# The Categories of personal data concerned

Clauses where the categories of personal data are comprehensively specified and not vague: **fully informative**

In other cases (e.g. when a clause only provides examples): **insufficiently informative**

Edreams Privacy Policy (last updated on 25 May 2018)

<cat2>If you sign up for our website using your social media account, link your account on our website to your social media account, or use certain other social media features of ours, we may access information about you via that social media provider in accordance with the provider's policies.  
</cat2>



# Substantive compliance

10 categories (GDPR art 5, 6, 9 and others)

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>

# Consent by using

(GDPR art 4(11)) (Rec 32)

When the consent is explicitly required: **fair processing clause**

Clauses stating that by simply using the service, the user consents to the terms of the privacy policy: **unfair processing clauses**

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

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



# Policy change

When notice is given and new consent is required: **fair processing clause**

When notice is given but a new consent (or confirmation of reading) is not required: **problematic processing clause**

Twitter Privacy Policy (effective on 25 May 2018)

**<pch2>**We may revise this Privacy Policy from time to time. The most current version of the policy will govern our processing of your personal data and will always be at <https://twitter.com/privacy>. If we make a change to this policy that, in our sole discretion, is material, we will notify you via an @Twitter update or email to the email address associated with your account.**</pch2>**



# Policy change

When no notice is given and new consent is not required: **unfair processing clause**

Booking Privacy Policy (last updated on 9 May 2018)

<pch3>We might amend the Privacy Statement from time to time. If you care about your privacy, visit this page regularly and you'll know exactly where you stand.</pch3>



# Clarity of expression

(GDPR art. 5(1)(a), 12(1) and others)

Is the privacy policy framed in an understandable and precise language?

4 main indicators of vagueness

Indicator	Language qualifiers
<b>1. Conditional Terms</b> The performance of a stated action or activity is dependent on a variable trigger	Depending, as necessary, as appropriate, as needed, otherwise reasonably, sometimes, from time to time, etc.
<b>Example</b>	<b>Rationale</b>
<vag> We also may share your information if we believe, in our sole discretion, that such disclosure is <b>necessary</b> :... "</vag>	The practice described as “necessary” suggests that the sharing will only occur in exceptional cases, however the clause fails to specify under what exceptional conditions the provider will disclose the information.

# Clarity of expression

Indicator	Language qualifiers
<b>2. Generalization:</b> i.e. terms that vaguely abstract information practices using contexts that are unclear. Action(s)/Information Types are vaguely abstracted with unclear conditions.	generally, mostly, widely, general, commonly, usually, normally, typically, largely, often, primarily, among other things, etc.
Example	Rationale
<pre>&lt;vag&gt; We <b>typically</b> or <b>generally</b> collect information ...&lt;/vag&gt; &lt;vag&gt; When you use an Application on a Device, we will collect and use information about you in <b>generally</b> similar ways and for similar purposes as when you use the TripAdvisor website.&lt;/vag&gt;</pre>	The use of the generalization term “generally” obscures for the data subject the service provider activities, since it provides a large flexibility to the service provider.

# Clarity of expression

(GDPR art. 5(1)(a), 12(1) and others)

Indicator	Language qualifiers
<p><b>3. Modality:</b> 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 ambiguous possibility of action or event.</p>	<p>may, might, could, would, possible, possibly, etc.</p>
Example	Rationale
<p>&lt;vag&gt;We <b>may</b> use your personal data to develop new services &lt;/vag&gt;</p>	<p>it is unclear whether or not the controller will use the data subject information to develop new services and in what cases and under</p>

# Clarity of expression

Indicator	Language qualifiers
<b>4. Non specific Numeric quantifiers:</b> which create ambiguity as to the actual measure	certain, numerous, some, most, many, various, including (but not limited to), variety
Example	Rationale
<code>&lt;vag&gt;When you create an Apple ID, apply for commercial credit, purchase a product, download a software update, register for a class at an Apple Retail Store, connect to our services, contact us or participate in an online survey, we may collect a <b>variety</b> of information, <b>including</b> your name, mailing address, phone number, email address, contact preferences, device identifiers, IP address, location information and credit card information.&lt;/vag&gt;</code>	it creates ambiguity with regard to the actual measure of information the data controller collect

# Clarity of expression

A combination of different forms of vagueness

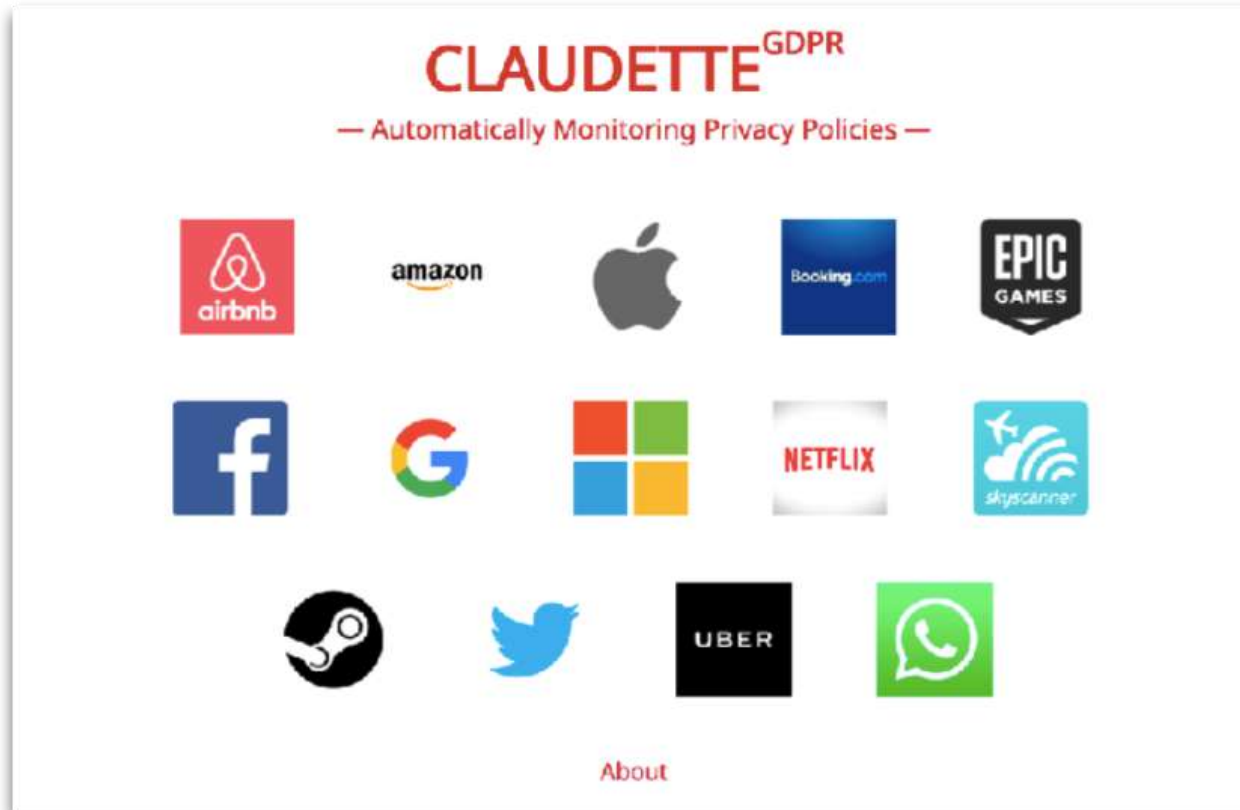
<vag>We **generally may** share personal information we collect with **certain** service providers, **some** of whom **may** use information for their own purposes **as necessary**.</vag>





# CLAUDETTE FOR GDPR

<http://claudette-gdpr.eu>



Manually annotated corpus

## WHERE DID WE START?

- 14 documents (32 now)
- 3,658 sentences
- 80,398 words
- 11.0% sentences contain unclear language
- 33.9% sentences contain potentially unlawful clauses

# CLAUDETTE FOR GDPR

http://claudette-gdpr.eu

## CLAUDETTE<sup>GDPR</sup>

— Automatically Monitoring Privacy Policies —

Use the checkboxes below to choose the sentences you want to highlight.  
In parenthesis, you can see the number of detected sentences for each category.  
By hovering your cursor over each unfair sentence, you can see the categories of each detected sentence.  
If you want to see the entire policy, click the grey [...] symbol; otherwise you will see just the detected clauses.

### Apple Privacy Policy

(153 sentences, 3643 words, 20861 characters)

Full Information (6)    Insufficient Information (19)    Unclear Language (12)    Problematic Processing (24)

[...]

You may be asked to provide your personal information anytime you are in contact with Apple or an Apple affiliated company .  
**Apple and its affiliates may share this personal information with each other and use it consistent with this Privacy Policy .**  
They may also combine it with other information to provide and  
**You are not required to provide the personal information that we need to provide our products or services or respond to any queries you may have**

Insufficient information clause for categories:  
•Categories of personal data concerned

Unclear language clause

[...]

When you share your content with family and friends using Apple products , send gift certificates and products , or invite others to participate in Apple services or forums , Apple may collect the information you provide about those people such as name , mailing address , email address , and phone number .  
**Apple will use such information to fulfill your requests , provide the relevant product or service , or for anti-fraud purposes .**  
In certain jurisdictions , we may ask for a government issued ID in limited circumstances including when setting up a wireless account and activating your device , for the purpose of extending commercial credit , managing reservations , or as required by law .

[...]

**The personal information we collect allows us to keep you posted on Apple 's latest product announcements , software updates , and upcoming events .**  
**If you do not want to be on our mailing list , you can opt out anytime by updating your preferences .**  
We also use personal information to help us create , develop , operate , deliver , and improve our products , services , content and advertising , and for loss prevention and anti-fraud purposes .

European University Institute  
DEPARTMENT OF LAW

claudette.eui.eu

■ 53

# Automated tagging prototype

<http://155.185.228.137/claurette4gdpr/>

**CLAUETTE**<sup>GDPR</sup>

Detecting Unclear Language in Privacy Policies (beta version)

Copy your text here

Submit

[About](#) [Cite](#) [Contact](#)

Accuracy: 68%

# Automated tagging prototype

<http://155.185.228.137/claurette4gdpr/>

## CLAUDETTE<sup>GDPR</sup>

Detecting Unclear Language in Privacy Policies (beta version)

Claudette found 56 clauses with unclear language (displayed in **bold**) out of 229 sentences.

[...]

**Learn more: Personal information we collect** **We collect personal information from you and any devices (including mobile devices) you use when you: use our Services, register for an account with us, provide us information on a web form, update or add information to your account, participate in our promotions, or contact us via email, phone, or discussion chat, or when you otherwise correspond with us.**

Unclear Language unfair clause

**Some of this personal information, such as a way to identify you, is necessary to enter into our User Agreement.**

**The provision of all other personal information is voluntary, but may be necessary in order to use our Services, such as the bidding, buying or selling information needed to conclude a transaction.**

**We may also collect personal information from other sources, as described below.**

**Personal information you give us when you use our Services or register for an account with us** Identifying information such as your name, addresses, telephone numbers or email addresses when you register for an account with us **Bidding, buying, or selling information you provide during a transaction, or other transaction-based content that you generate or that is connected to your account as a result of a transaction you are involved in** **Other content that you generate, or that is connected to your account (such as adding items to your basket, adding items to your Watch List, creating collections, and following other collections and sellers)** **Financial information (such as credit card or bank account numbers) in connection**

# Multilingualism: the German, Italian and Polish Claudette for ToS and PPs

 **CLAUDETTE** 

Ein automatisierter Detektor für potenziell unlautere Klauseln  
(GERMAN BETA VERSION)

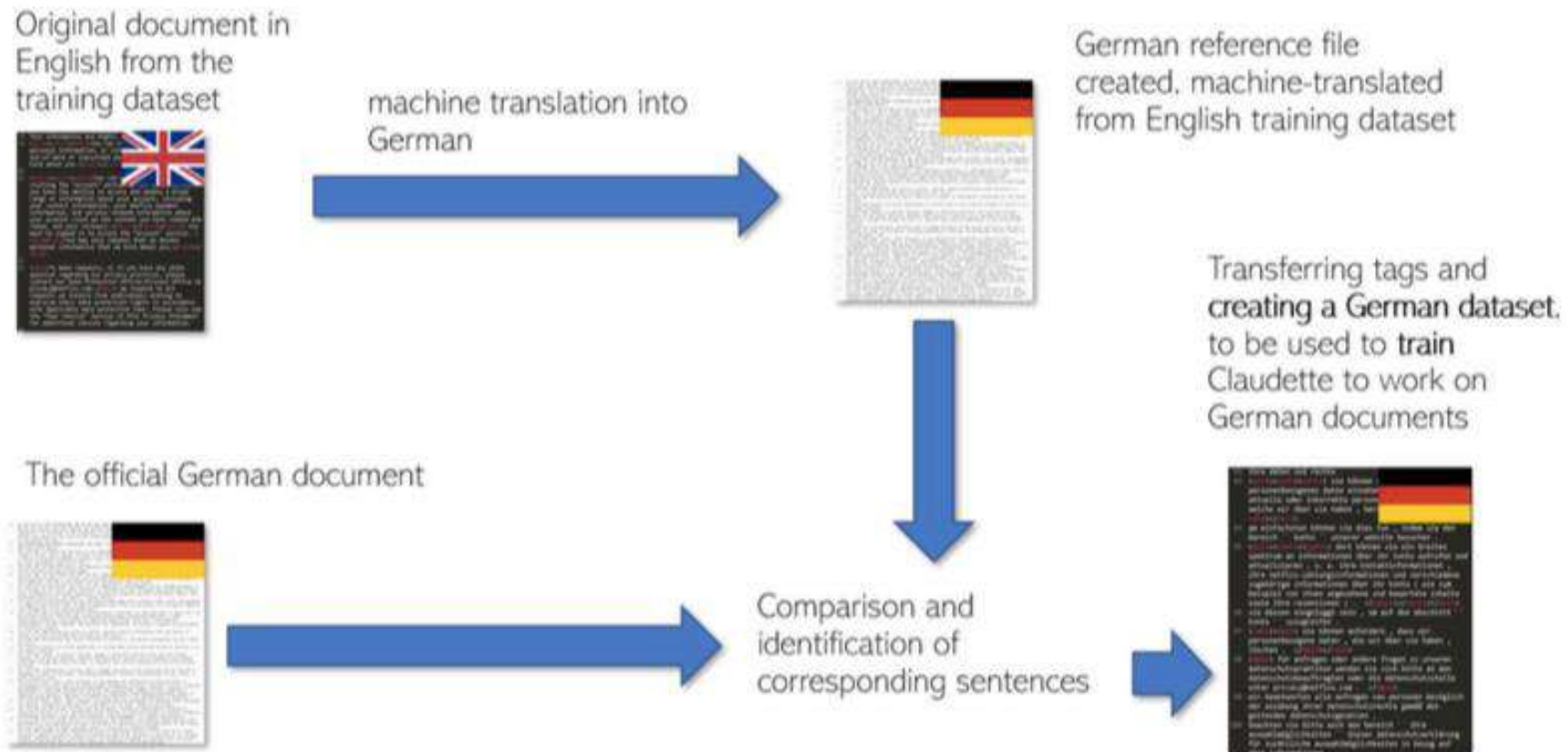
Kopieren Sie Ihren Text hier

Submit

[About](#) [Cite](#) [Contact](#)

# Multilingualism: the German, Polish and Italian Claudette

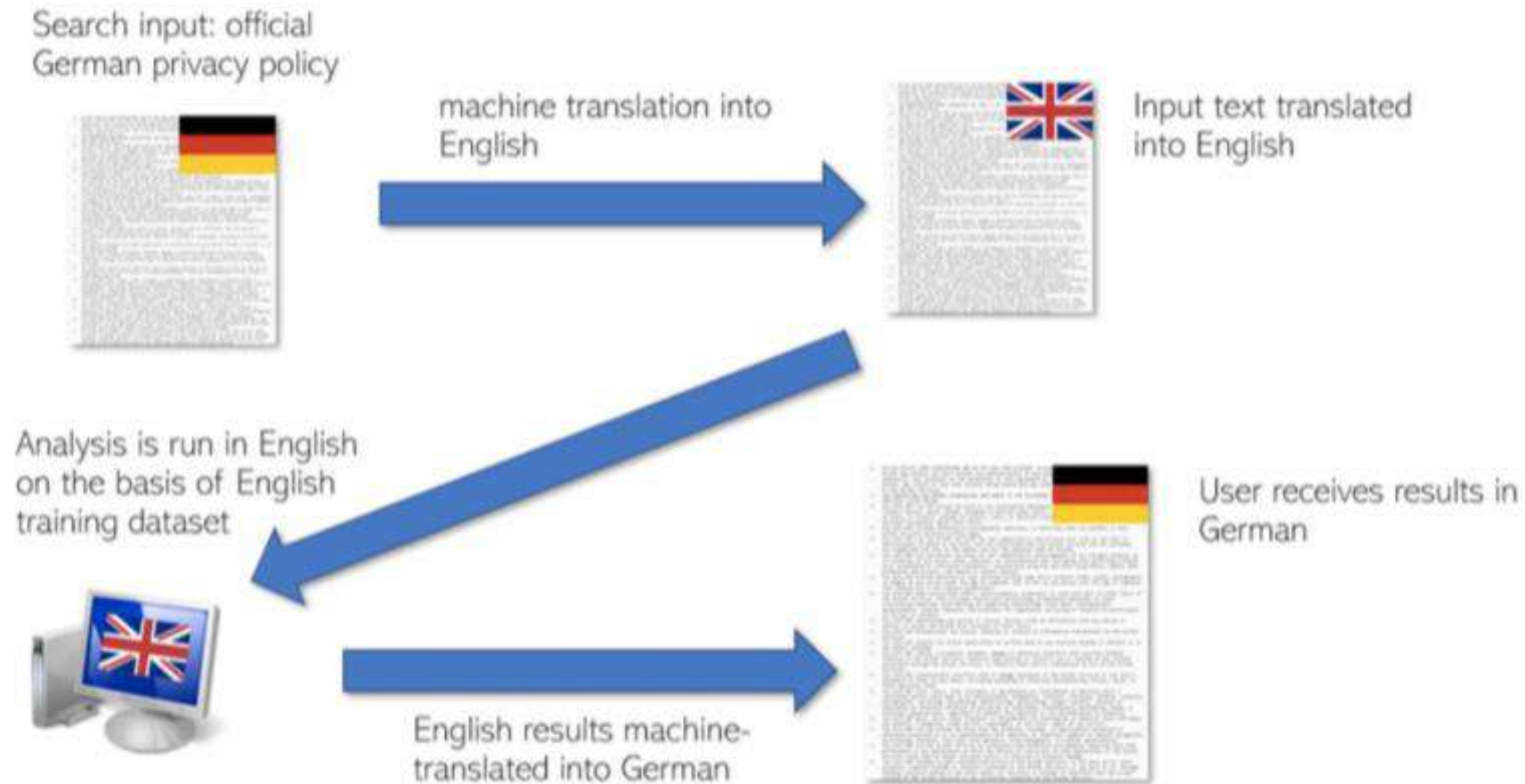
Approach 1: Semi-automated creation of a target-language dataset





# Multilingualism: the German Claudette

## Approach 2 – Machine translation of queries



# Multilingualism: the German Polish Italian Claudette

Approach 1 – Semi-automated creation of a target-language dataset

	DE	EN	IT	PL
a1	3	3	4	4
a2	22	29	29	35
a3	3	4	4	4
ch2	98	100	103	103
ch3	1	1	0	0
cr2	25	27	28	26
cr3	23	24	24	26
j1	14	15	15	15
j3	46	49	48	50
law1	18	16	16	19
law2	33	39	36	36
ltd1	19	27	16	17
ltd2	212	229	216	229
ltd3	1	1	1	1
pinc2	17	21	20	21
ter2	69	71	71	75
ter3	49	49	50	49
use2	54	58	58	61
total	707	753	739	771

**Dataset:** multilingual parallel corpus consisting of 25 Terms of Service annotated in English, Italian, German and Polish.

**Table 1: Corpus statistics.** we report the number of annotated clauses for each tag, across the four different languages. Suffices 1, 2, and 3 represent levels of fairness: 1 means clearly fair, 2 stays for potentially unfair, and finally 3 for clearly unfair.

claudette.eui.eu

# Multilingualism: the German Polish Italian Claudette

Approach 1 – Semi-automated creation of a target-language dataset

	DE	IT	PL
Precision	0.87	0.94	0.90
Recall	0.95	0.98	0.97
F1-macro	0.91	0.94	0.91
F1-micro	0.91	0.96	0.93
F1-weighted	0.91	0.96	0.93

Table 2: Projection results for the three languages.

# WEB-CRAWLER

Developed as a tool for automatic privacy policy monitoring

Two types of monitoring:

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

Earnings reports by e-mail

# Selected Publications

- LRuggeri F., Lagioia F., Lippi M., Torroni P., (2021) Detecting and explaining unfairness in consumer contracts through memory networks, in *Artificial Intelligence and Law*, Springer- Nature, 1-34
- Lippi, M.; Contissa, G.; Jablonowska, A.; Lagioia, F.; Micklitz, H-W; Palka, P.; Sartor, G.; Torroni, P., *The Force Awakens: Artificial Intelligence for Consumer Law*, *The journal of Artificial Intelligence Research*, 2020, 67., 169 - 190
- Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Sartor, G., & Torroni, P. CLAUDETTE: an Automated Detector of Potentially Unfair Clauses in Online Terms of Service, *Artificial Intelligence and Law*, Springer (2019).
- Lippi, M., Contissa, G., Lagioia, F., Micklitz, H. W., Palka, P., Sartor, G., & Torroni, P., Consumer protection requires artificial intelligence. *Nature Machine Intelligence*, 1, (2019).
- Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H.-W., Palka, P., Sartor, G., Torroni, P., CLAUDETTE meets GDPR: Automating the Evaluation of Privacy Policies using Artificial Intelligence Study Report, (BEUC), 2018.
- Lippi, M., Palka, P., Contissa, G., Lagioia, F., Micklitz, H. W., Sartor, G., & Torroni, P., Towards Consumer-Empowering Artificial Intelligence, JCAI-ECAI, Stockholm, special track on the evolution of contours of AI, (2018) .

# MAI4CAREU

Master programmes in Artificial  
Intelligence 4 Careers in Europe





## Bibliography

### 1. Science-Oriented AI

#### a. Mandatory:

- Castelfranchi, C., 2020, September. For a Science-oriented, Socially Responsible, and Self-aware AI: beyond ethical issues. In *2020 IEEE International Conference on Human-Machine Systems (ICHMS)* pp. 1-4. IEEE.
- Turing, A. M., 2022. "Computing Machinery and Intelligence." *Mind*, vol. 59, no. 236, 1950, pp. 433–60. *JSTOR*, <http://www.jstor.org/stable/2251299>. Accessed 19 July.
- Winfield, A.F., Michael, K., Pitt, J. and Evers, V., 2019. Machine ethics: The design and governance of ethical AI and autonomous systems [scanning the issue]. *Proceedings of the IEEE*, 107(3), pp.509-517.

#### b. Optional:

- Ricci, A., Tummolini, L. and Castelfranchi, C., 2019. Augmented societies with mirror worlds. *Ai & Society*, 34(4), pp.745-752.
- SHERPA Project, 2019. Briefings Released on the Ethical Implications of Smart Information Systems.

### 2. Introduction to Ethics

#### a. Mandatory:

- Floridi, L. and Cowls, J., 2022. A unified framework of five principles for AI in society. *Machine Learning and the City: Applications in Architecture and Urban Design*, pp.535-545.
- Shafer-Landau, R., 2020. The Kantian Perspective: Fairness and Justice. In *The fundamentals of ethics* (Fifth Edition). Oxford: Oxford University Press.
- Shafer-Landau, R., 2020. Virtue Ethics. In *The fundamentals of ethics* (Fifth Edition). Oxford: Oxford University Press.
- Shafer-Landau, R., 2021, What is Morality? *Living Ethics. An Introduction with Readings*, II Edition, Oxford: Oxford University Press.
- Shafer-Landau, R., 2021, Consequentialism *Living Ethics. An Introduction with Readings*, II Edition, Oxford: Oxford University Press.

#### b. Optional:

- Allen, C., Smit, I. and Wallach, W., 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, 7(3), pp.149-155.
- Bostrom, N. and Yudkowsky, E., 2018. The ethics of artificial intelligence. In *Artificial intelligence safety and security*, pp. 57-69. Chapman and Hall/CRC.
- Taddeo, M., 2016. The moral value of information and information ethics. In *The Routledge Handbook of Philosophy of Information*, pp. 377-390. Routledge.

- Shafer-Landau, R., 2020. The Social Contract Tradition. In *The fundamentals of ethics* (Fifth Edition). Oxford: Oxford University Press.

### 3. Do Artifacts have politics?

#### a. Mandatory

- Winner, L., 1980. Do artifacts have politics?. In *Modern Technology: Problem or Opportunity?*, (109, 1). MIT Press. pp. 121-136

#### b. Optional:

- Contissa, G., 2017. Automation and Liability: an analysis in the context of socio-technical systems. In *i-lex*, (11, 1) pp. 15-45.
- Vermaas, P., Kroes, P., van de Poel, I., Franssen, M., Houkes, W. 2011. Ethics and Designing. In: *A Philosophy of Technology. Synthesis Lectures on Engineers, Technology, & Society*. Springer, Cham.

### 4. Introduction to Ethics and Digital Humanism

#### a. Mandatory:

- Floridi, L. and Taddeo, M., 2016. What is data ethics? *Philosophical Transactions of the Royal Society A*, 374.

### 5. Value Alignment

#### a. Mandatory:

- Russell, S., Dewey, D. and Tegmark, M., 2015. Research priorities for robust and beneficial artificial intelligence. *Ai Magazine*, 36(4), pp.105-114.

#### b. Optional:

- Rossi, F. and Mattei, N., 2019, July. Building ethically bounded AI. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(1), pp. 9785-9789.
- Loreggia, A., Mattei, N., Rossi, F. and Venable, K.B., 2018, July. A notion of distance between cp-nets. In *Proceedings of AAMAS*, pp. 955-963.
- Burton, E., Goldsmith, J. and Mattei, N., 2018. How to teach computer ethics through science fiction. *Communications of the ACM*, 61(8), pp.54-64.
- Rossi, F. and Loreggia, A., 2019, May. Preferences and Ethical Priorities: Thinking Fast and Slow in AI. In *Proceedings of AAMAS*, pp. 3-4.

### 6. AI and Human Rights

#### a. Mandatory:

- Floridi, L., Cows, J., Beltrametti, M. et al., 2018. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds & Machines*, 28, pp. 689–707.
- European Digital Rights (EDRi), 2021. Open letter: Civil society call for the introduction of red lines in the upcoming European Commission proposal on Artificial Intelligence.

- Council of Europe Commissioner for Human Rights, 2019. Unboxing artificial intelligence: 10 steps to protect human rights.
- Sartor, G., 2017. Human rights and information technologies. *The Oxford Handbook of Law, Regulation and Technology*, pp.424-458.

b. Optional:

- Sen, A., 2017. Elements of a theory of human rights. In *Justice and the capabilities approach*, pp. 221-262. Routledge.
- Internet Rights & Principles Coalition, 2014. The charter of human rights and principles for the internet. Internet Governance Forum, United Nations.

## 7. Logic Programming

a. Mandatory:

b. Optional:

- Apt, K. R. 2005., The logic programming paradigm and Prolog. In Mitchell, J. C., editor, *Concepts in Programming Languages*, chapter 15, pp. 475–508. Cambridge University Press, Cambridge, UK.
- Baroni, P. and Giacomin, M. 2007. On principle-based evaluation of extension-based argumentation semantics. *Artificial Intelligence*, 171(10) pp.675–700.
- Baroni, P. and Giacomin, M., 2009. Semantics of Abstract Argument Systems, pp. 25–44. Springer US, Boston, MA.
- Borning, A., Maher, M. J., Martindale, A., and Wilson, M., 1989. Constraint hierarchies and logic programming. In Levi, G. and Martelli, M., editors, 6th *International Conference on Logic Programming*, volume 89, pp. 149–164, Lisbon, Portugal. MIT Press.
- Calegari, R., Ciatto, G., Denti, E., and Omicini, A. 2020. Logic-based technologies for intelligent systems: State of the art and perspectives. *Information*, 11(3), pp.1–29.
- Calegari, R., Denti, E., Dovier, A., and Omicini, A. 2018a. Extending logic programming with labelled variables: Model and semantics. *Fundamenta Informaticae*, 161(1-2), pp.53–74.
- Calegari, R., Denti, E., Mariani, S., and Omicini, A. 2018b. Logic programming as a service. *Theory and Practice of Logic Programming*, 18(3-4), pp.1–28.
- Dung, P. M. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2), pp.321–357.
- Dyckhoff, R., Herre, H., and Schroeder-Heister, P., 1996. Extensions of Logic Programming, *5th International Workshop, ELP'96*, volume 1050 of LNCS, Leipzig, Germany. Springer.
- Levesque, H. J. 1989. A knowledge-level account of abduction. In *IJCAI*, pp. 1061–1067.
- Omicini, A., Ricci, A., and Viroli, M. 2008. Artifacts in the A&A meta-model for multi-agent systems. *Autonomous Agents and Multi-Agent Systems*, 17(3), pp.432–456.
- Poole, D. 1993. Logic programming, abduction and probability. *New Generation Computing*, 11(3–4), p.377.

- Riveret, R., Oren, N., and Sartor, G. 2020. A probabilistic deontic argumentation framework. *International Journal of Approximate Reasoning*, 126, pp.249–271.
- Saptawijaya, A. and Pereira, L. M. 2019. From logic programming to machine ethics. In Bendel, O., editor, *Handbuch Maschinenethik*, pp. 209–227. Springer VS, Wiesbaden.

## 8. Data Protection

### a. Mandatory:

- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)
- Sartor, G., and Lagioia, F., 2020. Study: The impact of the General Data Protection Regulation on artificial intelligence. European Parliament.

### b. Optional:

- Zarsky, T.Z., 2016. Incompatible: The GDPR in the age of big data. *Seton Hall L. Rev.*, 47, pp.995-1020.

## 9. A Framework for Ethical Principles

### a. Mandatory:

- Tsamados, A., Aggarwal, N., Cows, J., Morley, J., Roberts, H., Taddeo, M., and Floridi, L., 2021. The ethics of algorithms: Key problems and solutions. In Floridi, L., editor, *Ethics, Governance, and Policies in Artificial Intelligence*. Springer.

## 10. Fairness in Automated Decisions

### a. Mandatory:

- Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A., 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1), pp. 3-44.
- Lagioia, F., Rovatti, R., & Sartor, G., 2022. Algorithmic fairness through group parities? The case of COMPAS-SAPMOC. *AI & SOCIETY*, pp. 1-20.

## 11. Autonomous Vehicles

### a. Mandatory:

- Contissa, G., Lagioia, F., & Sartor, G., 2017. The Ethical Knob: ethically-customisable automated vehicles and the law. *Artificial Intelligence and Law*, 25(3), 365-378.
- Nyholm, S., 2018. The ethics of crashes with self-driving cars: A roadmap, II. *Philosophy Compass*, 13(7).
- Nyholm, S., 2018. The ethics of crashes with self-driving cars: A roadmap, I. *Philosophy Compass*, 13(7).

### b. Optional:

- Etzioni, A., & Etzioni, O., 2017. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21(4), pp. 403-418.
- Gentzel, M., 2020. Classical liberalism, discrimination, and the problem of autonomous cars. *Science and Engineering Ethics*, 26(2), pp. 931-946.
- Nyholm, S., 2018. Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, 24(4), pp. 1201-1219.

## 12. Intelligent Weapons

### a. Mandatory:

- House, P. B., 2015. Autonomous weapons systems: five key human rights issues for consideration. *Amnesty International Publications. London, 28*.
- Arkin, R., 2018. Lethal autonomous systems and the plight of the non-combatant. In *The political economy of robots*, pp. 317-326. Palgrave Macmillan, Cham.
- Arkin, R., et al., 2019. Autonomous Weapon Systems: A Roadmapping Exercise, *Policy workshop organized by Max Tegmark, Emilia Javorsky and Meia Chita-Tegmark*.
- Scharre, P., Horowitz, M. C., 2015. An Introduction to AUTONOMY in WEAPON SYSTEMS, *CNAS Working Papers*.
- Walzer, M., 2015. Just and unjust wars: A moral argument with historical illustrations. Hachette UK, chapters 8,9.

### b. Optional:

- Arkin, R., 2009. Governing lethal behavior in autonomous robots. Chapman and Hall/CRC, chapter 10.
- Sartor, G., & Omicini, A., 2016. The autonomy of technological systems and responsibilities for their use. In N. Bhuta, S. Beck, R. Geiß, H. Liu, & C. Kreß (Eds.), *Autonomous Weapons Systems: Law, Ethics, Policy* (pp. 39-74). Cambridge: Cambridge University Press.
- Sharkey, N., 2014. Towards a principle for the human supervisory control of robot weapons. *Politica & societa*, 3(2), pp. 305-324.
- Sharkey, A., 2019. Autonomous weapons systems, killer robots and human dignity. *Ethics and Information Technology*, 21(2), pp. 75-87.

## 13. Ethics of Filtering

### a. Mandatory:

- Sartor, G., and Loreggia, A., Study: The impact of algorithms for online content filtering or moderation (“upload filters”). European Parliament, 2020.

## 14. AI and Unfairness in ToS and PPs

### a. Mandatory:

- Drawzeski, K., Galassi, A., Jablonowska, A., Lagioia, F., Lippi, M., Micklitz, H.W., Sartor, G., Tagiuri, G. and Torroni, P., 2021, November. A Corpus for Multilingual Analysis of Online Terms of Service. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 1-8.

- Contissa, G., Docter, K., Lagioia, F., Lippi, M., Micklitz, H.W., Pałka, P., Sartor, G. and Torroni, P., 2018. Claudette meets gdpr: Automating the evaluation of privacy policies using artificial intelligence. Study Report, Funded by The European Consumer Organisation (BEUC).
- Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.W., Sartor, G. and Torroni, P., 2019. CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, 27(2), pp.117-139.
- Ruggeri, F., Lagioia, F., Lippi, M. and Torroni, P., 2022. Detecting and explaining unfairness in consumer contracts through memory networks. *Artificial Intelligence and Law*, 30(1), pp.59-92.
- Lippi, M., Contissa, G., Jablonowska, A., Lagioia, F., Micklitz, H.W., Palka, P., Sartor, G. and Torroni, P., 2020. The force awakens: artificial intelligence for consumer law. *Journal of artificial intelligence research*, 67, pp.169-190.

b. Optional:

- Lagioia, F., Jabłonowska, A., Liepina, R. and Drazewski, K., 2022. AI in Search of Unfairness in Consumer Contracts: The Terms of Service Landscape. *Journal of Consumer Policy*, pp.1-56.