DataAI TPT 951 AI Ethics

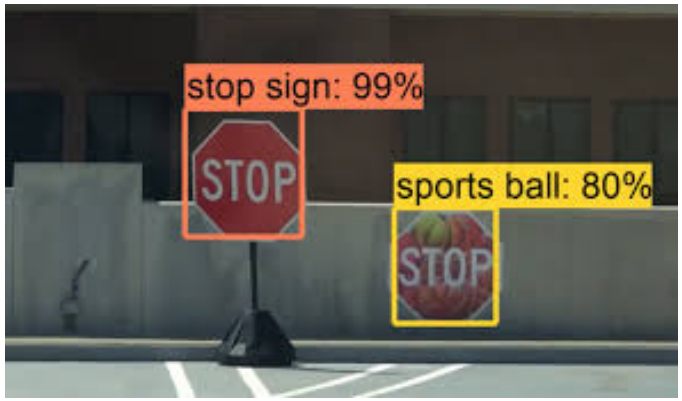
Winston Maxwell

Telecom Paris, Institut Polytechnique de Paris

telecom-paris.fr/ai-ethics

November 24, 2020

AI needs to grow up…



Artificial intelligence Oct 25

A biased medical algorithm favored white people for health-care programs

Self-driving Uber car that hit and killed woman did not recognize that pedestrians jaywalk

The automated car lacked "the capability to classify an object as a pedestrian unless that object was near a crosswalk," an NTSB report said.

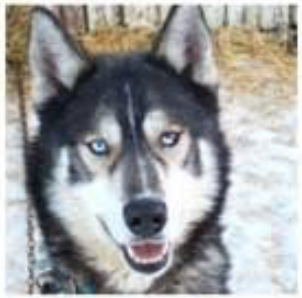
Google's solution to accidental algorithmic racism: ban gorillas

Google's 'immediate action' over AI labelling of black people as gorillas was simply to block the word, along with chimpanzee and monkey, reports suggest

Artificial intelligence / Machine learning

AI could help with the next pandemic—but not with this one

Some things need to change if we want AI to be useful next time, and you might not like them.



(a) Husky classified as wolf

(b) Explanation

Course schedule

Nov 24 and Dec 1, 2020:

AI Ethics, law and fundamental rights

Winston Maxwell

Dec 8 and Dec 15, 2020:

Algorithmic bias and discrimination

Sophie Chabridon

Jan 5 and Jan 12, 2021:

Privacy and data security, plus hackathon

Fabian Suchanek

Jan 19, 2021:

Social impacts of AI

Ada Diaconescu

Jan 26, 2021:

Presentation of student posters
Winston Maxwell

Feb 2, 2021:

Final exam

All classes are from 13:30 to 16:45 by Zoom video only

Grades

Final exam: 50%

Hackathon (Prof. Suchanek): 10%
Social impacts class participation (Prof. Diaconescu): 10%
Bias and discrimination class participation (Prof. Chabridon): 10%
Individual posters: 20%

The posters project

- Each of you will pick a **use case or anecdote** to present in a poster that highlights an AI ethics question.
- In the poster you :
  - Present the artefact illustrating the use case or anecdote (can be a film, a book, an article, TV series, academic paper, MIT Technology Review article…)
  - What is the ethical problem raised? Link the problem to one of the « over-arching AI ethics themes » of the course or one of the Asilomar AI principles
  - How should the ethical problem be approached: elements of analysis, arguments for and against.
  - Link the problem to a legal text (eg an article in the EU Charter of Fundamental Rights)
  - Proposed solution

Over-arching themes of AI Ethics (1/2)

I. AI and the effect on work

a. AI replacing workforce

b. What is the role of work in human existence?

c. AI for recruiting

d. Amazon Mechanical Turk

II. AI and the surveillance state

a. **"Surveillance capitalism"** (Shoshana Zuboff) – transforming ubiquitous surveillance and data gathering into business opportunities (Facebook and Google)

b. **Surveillance by government**: predictive policing, facial recognition, algorithms to detect terrorist threats: how to draw the right balance between privacy and public security

III. AI and health

a. Individualized, predictive medicine

b. Epidemic (COVID) management

c. Neuralink

d. Augmented humans, transhumanism

e. Robot doctors

IV. AI and democratic institutions

a. AI and manipulation of populations

b. Fake news, polarization, and the "post-truth" era

c. Election manipulation, social (cyber) warfare

d. Freedom of expression vs censorship

Over-arching themes of AI Ethics (2/2)

V. AI and human dignity

a. Autonomous lethal weapons: the respective role of humans and machines in warfare

b. Robot judges – can humans be judged by a machine? (cf. Estonia robot judges experiment)

VI. AI and discrimination

a. Racism, gender inequality, social inequalities. Does AI make societal discriminations worse? Can AI help offset human discriminations?

VII. AI and the end of serendipity

a. What is the role of chance in our lives, careers, scientific discoveries? By reducing the role of chance, does AI harm innovation and personal development?

b. Can chance be a justifiable solution for ethical dilemmas such as the trolley problem? (Alexei Grinbaum)

c. The role of outliers ("black swans") in human development.

VIII. AI and human psychology

a. Human machine interactions - how can AI make humans smarter (and not dumber)

b. Robot companions, robot 'emotions'

c. Social engineering - nudges to help affect human behavior: eg « you haven't been walking enough today… »

IX. AI and safety certification

How does machine learning change our approach to certifying safety-critical systems? What AI-related safety lessons can we learn from the Boeing 737 Max failures?

X. Can AI save humanity from itself?

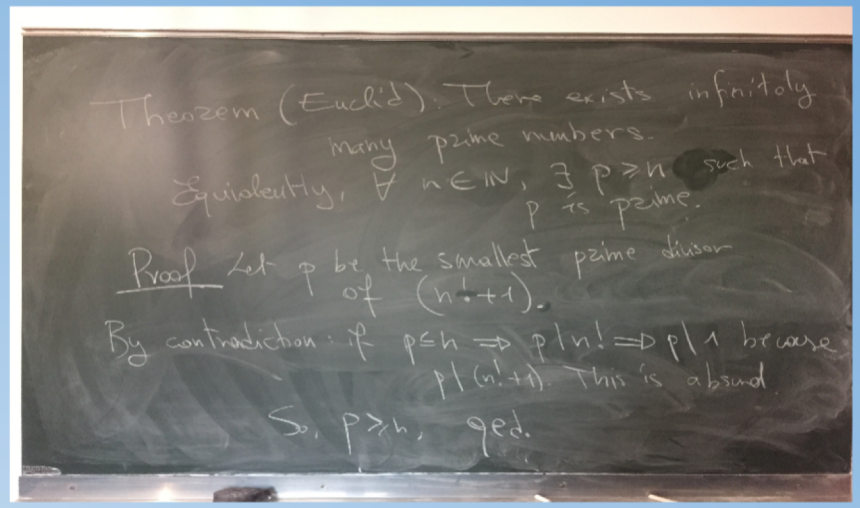
a. AI and climate change

b. AI "taking control": Isaac Asimov laws of robotics, 2001 Space Odyssey, etc.

The end of research in pure mathematics?

Abstract: Until now research in pure mathematics has been done with chalks and blackboard. The only software used by researchers was a text editor. The validity of a result was checked by peers only and this led to many wrong papers. Since 2013 the efficient proof-checking software "Lean" has been developed and some renown mathematicians are actively working with it.

Problem: Will in a near future a proof of a theorem be considered correct only if approved by "Lean"? Will in a longer term "Lean" replace mathematicians, being able to independently discover new theorems?




```
321
322  /-- Euclid's theorem. There exist infinitely many prime numbers.
323  Here given in the form: for every `n`, there exists a prime number `p ≥ n`. -/
324  theorem exists_infinite_primes (n : ℕ) : ∃ p, n ≤ p ∧ prime p :=
325  let p := min_fac (n! + 1) in        Anatole Dedecker, a month ago • feat(data/nat
326  have f1 : n! + 1 ≠ 1, from ne_of_gt $ succ_lt_succ $ factorial_pos _,
327  have pp : prime p, from min_fac_prime f1,
328  have np : n ≤ p, from le_of_not_ge $ λ h,
329    have h₁ : p ∣ n!, from dvd_factorial (min_fac_pos _) h,
330    have h₂ : p ∣ 1, from (nat.dvd_add_iff_right h₁).2 (min_fac_dvd _),
331    pp.not_dvd_one h₂,
332  ⟨p, np, pp⟩
333
```

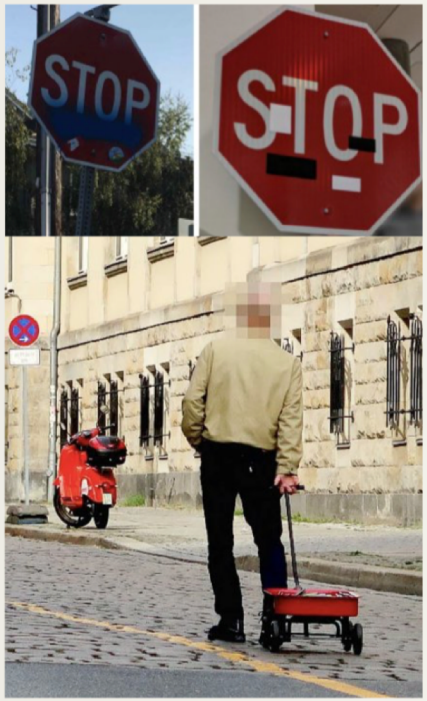
Analysis: The mathematical community is currently sceptical about "Lean". Most mathematicians do not want to spend time learning a new software and they fear that "Lean" will steal their job. An efficient tool which could help their work is being boycotted.

Valentina Di Proietto

Self-driving vehicle, insurance & legislation: who is or would be liable in case of an accident?

Fact: in March 18th 2018 Arizona (USA) an Uber's self-driving car did not stop and the backup driver on board did not brake killing a pedestrian (https://www.nytimes.com/2020/09/15/technology/uber-autonomous-crash-driver-charged.html)

Today: stakeholders, **ethical** promises and risks or obstacles, technologies, SAE levels, legislation (Vienna convention) and deployment roadmap



High-risk AI: Safety? (Cyber-)security? **Liability**? Privacy? Transparency? Affordability?

50% of the total car cost dedicated to electronics and software by 2030

Car and insurance industries **disruptions**? (e.g. Tesla insurance)

In France:
2014: first experiments
2020: SAE level-3 automated cars on roads
2022: SAE level-4

SAE lvl 0: no automation
SAE level 4: the car can *"intervene if the driver does not respond **appropriately**"*
SAE level 5: fully autonomous

Tomorrow: will a fully autonomous vehicle be authorized on the roads? opportunities, upcoming technical challenges, and ethical considerations

Adele Harrissart (MS-IA)

Introduction to ethical traditions

Virtue ethics: Aristotle: a model for iterative learning and growth, and moral value informed by context and practice, not just as compliance with a given, static ruleset.

Deontological ethics: Immanuel Kant.

> "Act only on that maxim through which you can at the same time will that it should become a universal law."
>
> "Act in such a way that you always treat humanity, whether in your own person or in the person of any other, never simply as a means, but always at the same time as an end."

Utilitarian ethics: According to the utility principle, the right course of action is the one that maximizes the utility (utilitarianism) or pleasure (hedonism) for the greatest number of people. (cf. Jeremy Bentham)

Ethics of care: Relationships are ontologically basic to humanity,; to care for other human beings is one of our basic human attributes.

Ubuntu: Its basic tenet is that a person is a person through other persons.

Shinto tradition is an animistic religious tradition, positing that everything is created with, and maintains, its own spirit (*kami*) and is animated by that spirit.

(source: The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, chapter on classical ethics)

Ethics and law: an overlap?

Ethics and law

- You shall not kill
- You shall not lie (in commerce)
- You shall help someone in danger
- You shall respect human dignity

Ethics but not law

- You shall not lie (to your friends and family)
- You shall help someone who needs money

What about sanctions?

Potential conflicts between ethics and law?

Fundamental rights

CHARTER OF FUNDAMENTAL RIGHTS OF THE EUROPEAN UNION

(2012/C 326/02)

- Human dignity
- Right to life
- Right to integrity of the person
- Prohibition of torture
- Prohibition of slavery
- Liberty and security
- Privacy and protection of personal data
- Freedom of religion and association
- Freedom of expression and access to information

Text of European Convention on Human Rights

- Right to a fair trial
- Right to an effective remedy
- Right to marry
- Right to property
- Right to conduct a business
- Prohibition of discrimination

Text of EU Charter of Fundamental Rights

Some rights are **absolute**

Others are not, and can be **balanced** against other rights.

- The balancing is called the **proportionality test**.

The role of fundamental rights

Fundamental rights are at the top of the legal food chain

- EU Treaty, Charter, European Convention on Human rights, constitutions
- EU directives and regulations
- National statutes
- Regulations
- Soft law

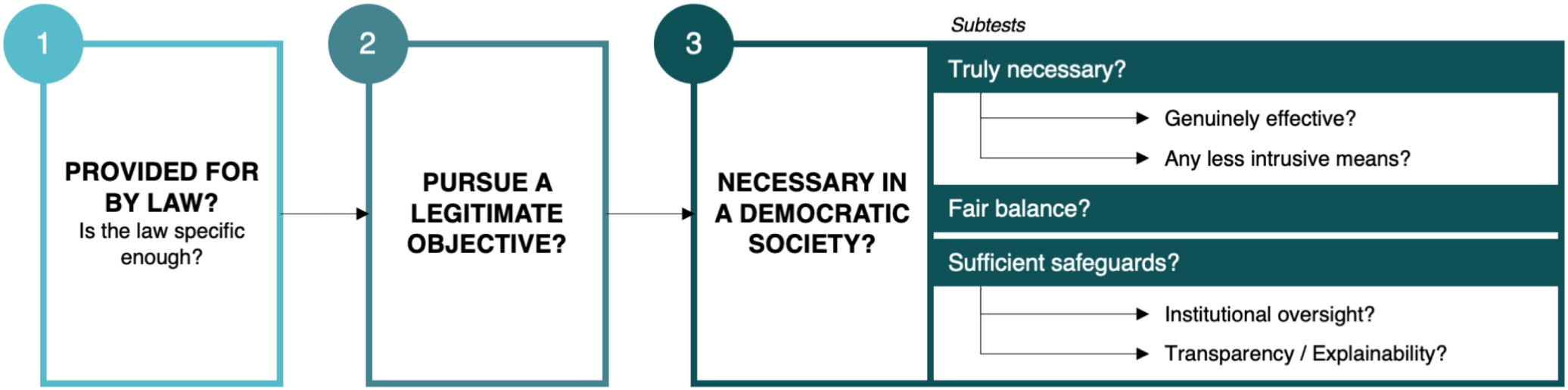
Fundamental rights' vertical effect

- Fundamental rights protect the citizens against the state
- But sometimes they protect againt actions by private entities

The proportionality test

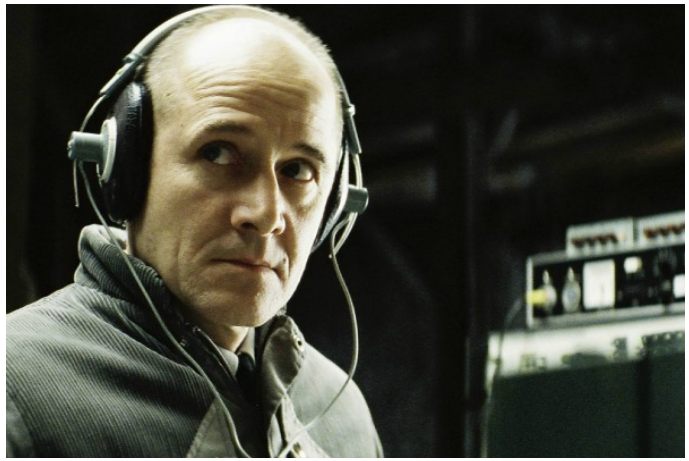
- Applies when rights come into conflict

The proportionality test



Example of state surveillance

A machine learning tool to detect potential terrorist activity from communications traffic and location data



Fundamental rights at stake:

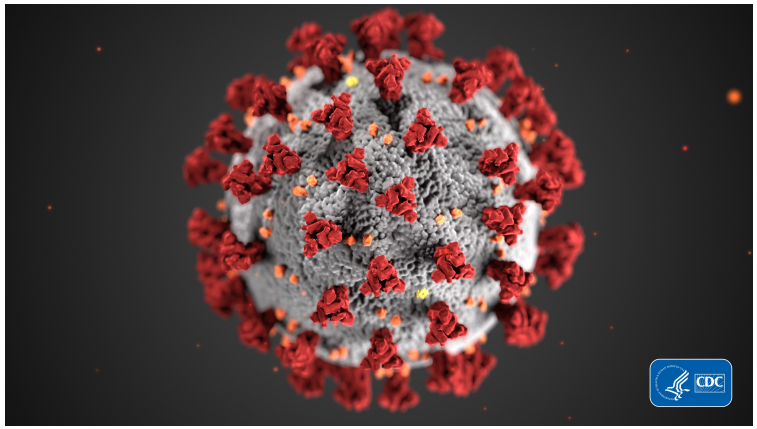
- Right to public security (protect against terrorist attacks)
- Right to protection of privacy
- Right to freedom of expression

Conditions imposed by the European Court of Justice (6 Oct. 2020):

- Provided for by law? Yes
- Legitimate objective? Yes
- Necessary in a democratic society?
- only during limited times where high terrorist threat
- institutional oversight
- testing, and human supervision

Example of fighting COVID

Using phone data to track infected individuals and potential contacts



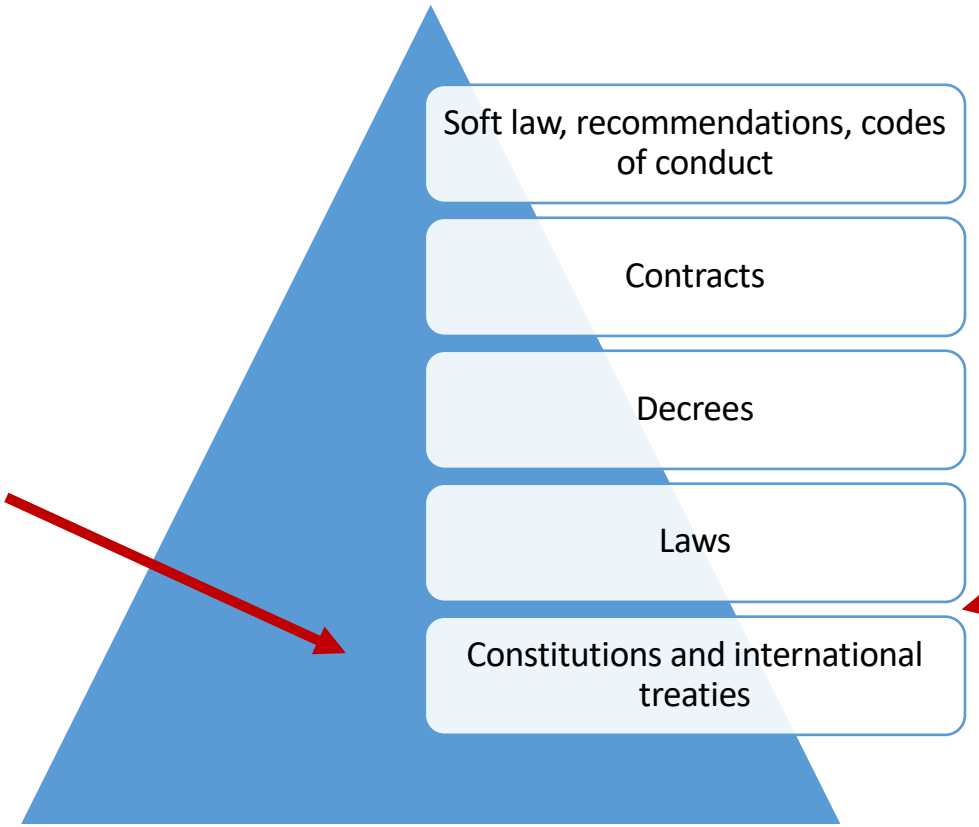
Fundamental rights at stake:

- Right to public health
- Right to protection of privacy
- Right to freedom of association

Let’s go through the proportionality test!

Law is built in layers (like the internet!)

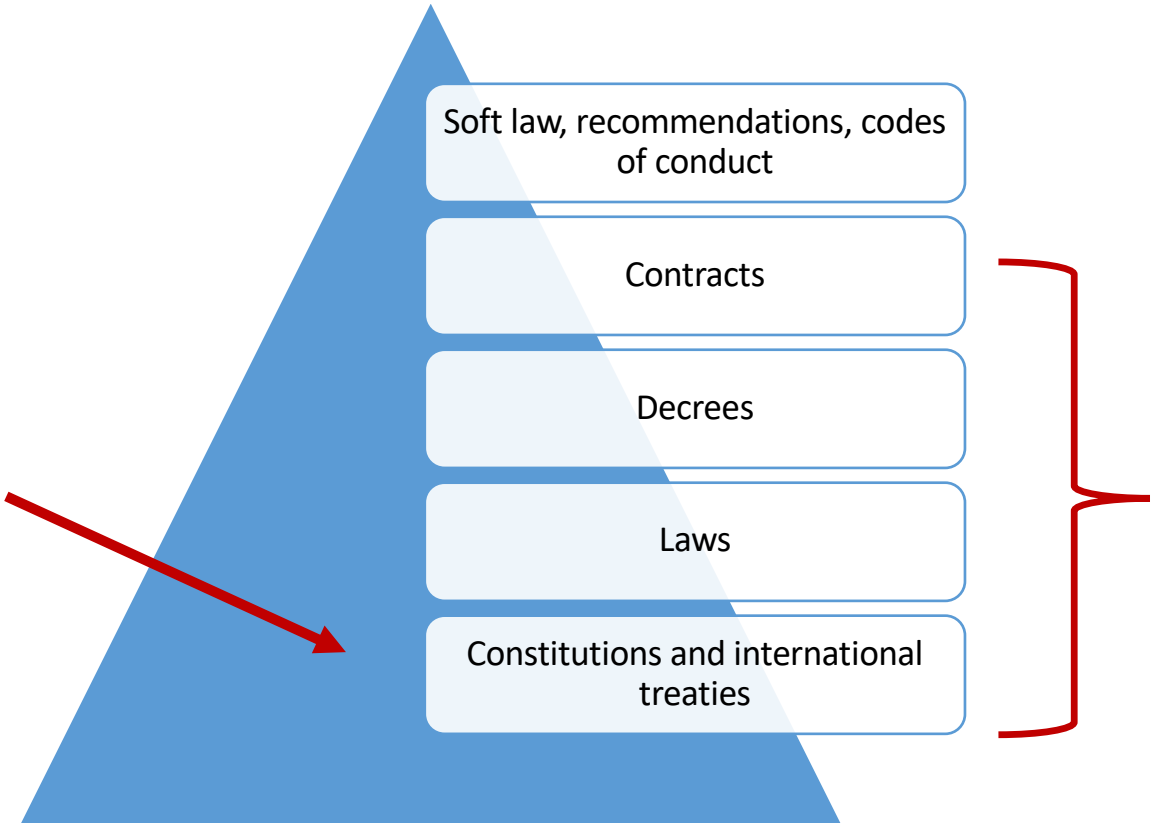
Constitutions (eg. Déclaration des droits de l'homme et du citoyen de 1789) And international treaties (eg.European convention on human rights) make up the foundational layer.



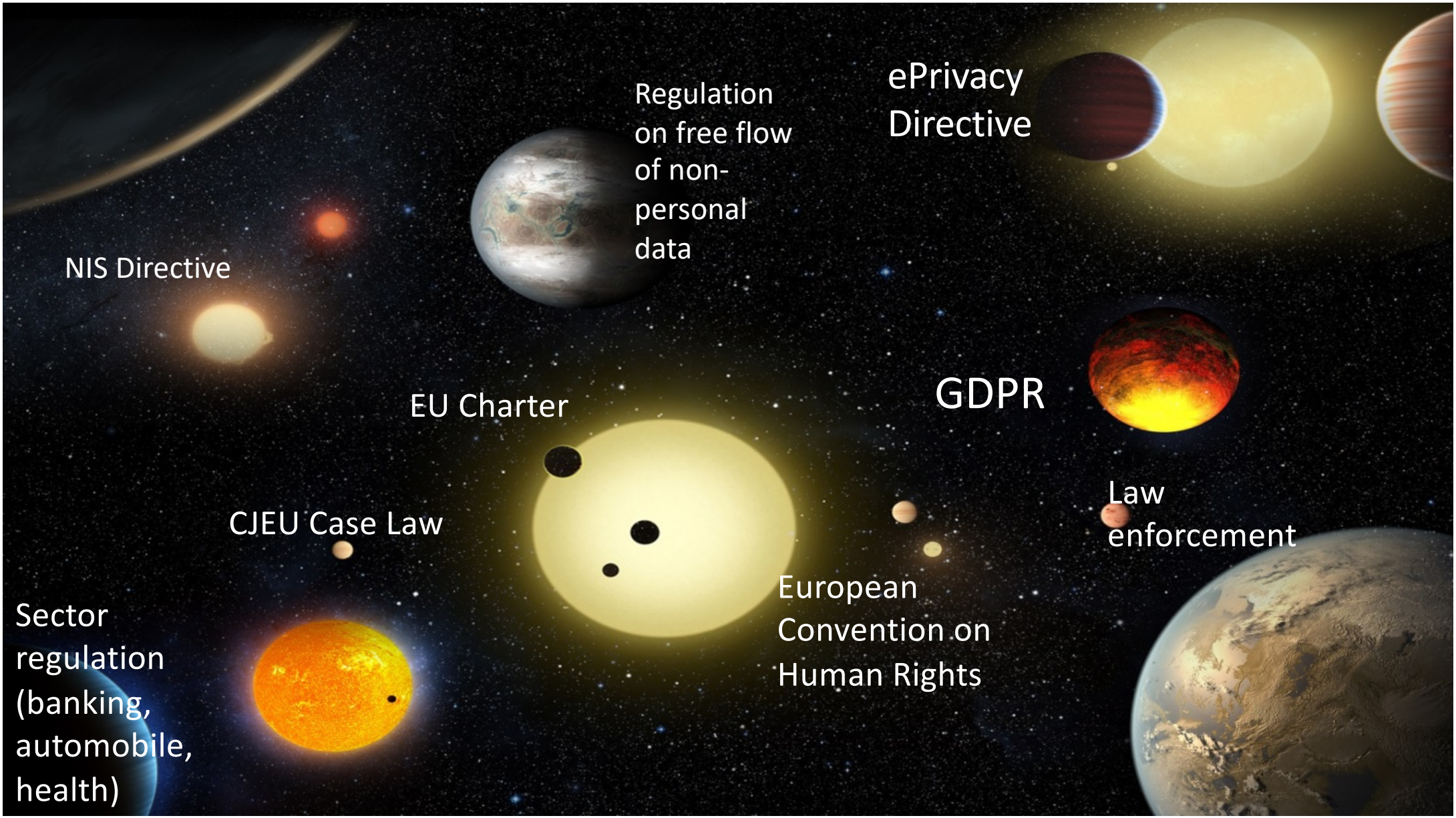
EU Directives and regulations

Law is built in layers (like the internet!)

Constitutions (eg. Déclaration des droits de l'homme et du citoyen de 1789) And international treaties (eg.European convention on human rights) make up the foundational layer.



Sanctions through courts and police


Regulation on free flow of non-personal data
ePrivacy Directive
NIS Directive
GDPR
EU Charter
CJEU Case Law
Law enforcement
European Convention on Human Rights
Sector regulation (banking, automobile, health)

Without enforcement, law remains a dead letter

INTERNAL GOVERNANCE

- AI ethics committee
- Chief Ethics Officer
- Data Protection Officer
- Internal Audit
- Data Management Officer

EXTERNAL OVERSIGHT

- Regulatory authorities (CNIL, ACPR, AMF, CSA, ARCEP...)
- Self-regulatory organizations
- External audit

LAWS AND REGULATIONS

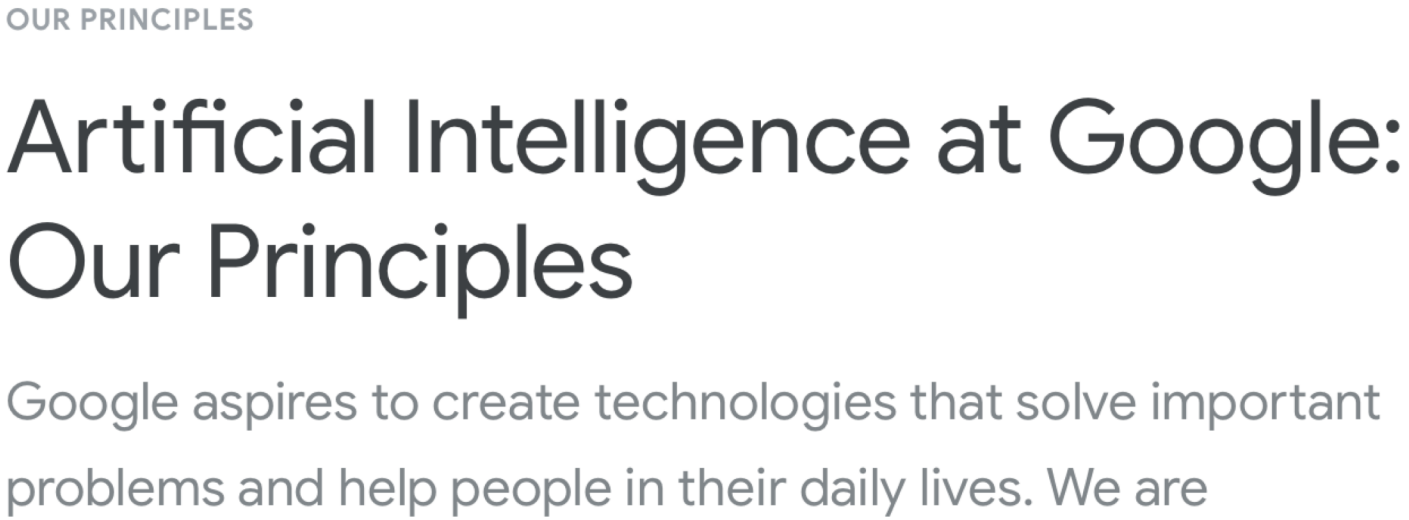
- General statutes (Civil Code, GDPR...)
- Sector-specific laws (Health Code, Transport Code...)
- AI-specific laws

COURT ENFORCEMENT

- Speed of court decisions
- Territoriality and jurisdiction
- Access to technical information

Question: how are ethics enforced?

Are ethics charters useful?



OUR PRINCIPLES

Artificial Intelligence at Google: Our Principles

Google aspires to create technologies that solve important problems and help people in their daily lives. We are

- Why do groups create ethics charters?

- What is their social value?

  See Karen Yeung et al. Human Rights-Centred Design, Deliberation and Oversight: An End to Ethics Washing.

The trolley problem

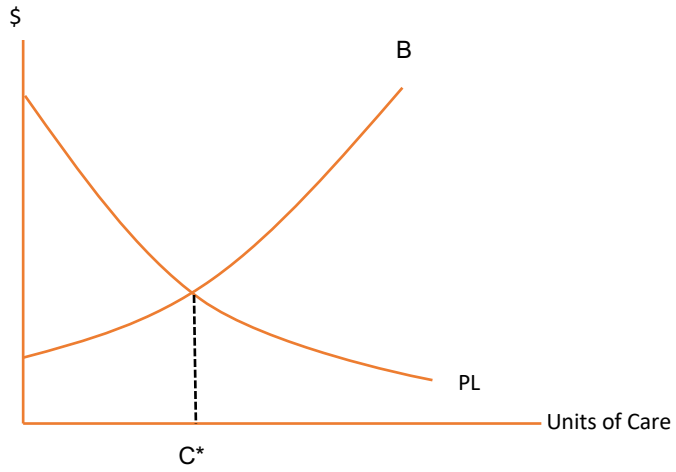
https://www.youtube.com/watch?v=yg16u_bzjPE

Who is liable for this?

What can we learn from the Boeing 737 Max crashes?

Principles of liability

Fault-based liability: a person is liable only if he or she failed to exercise a **reasonable level of care.**
Hand formula: a person is negligent if he or she expends costs on injury prevention ("B") in an amount less than the amount of the injury "L" multiplied by its probability "P". When calculating "P", the injuring party can assume that the victim will also take reasonable steps to avoid injury. Under this approach, not all injuries are prevented, only a reasonable level of injuries. In its simplest form, the Hand formula means that person will be negligent if, but only if, $B < PL$.



The ‘least cost avoider'

Other approaches to liability

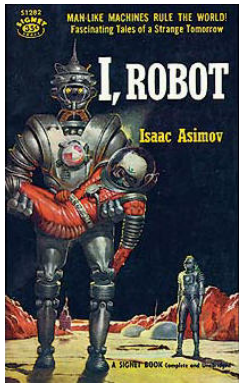
Product liability: the manufacturer is liable for damage caused by any product that is ‘unsafe’.

Strict liability: a person who pursues a dangerous activity is strictly liable for the harm caused.

Mandatory insurance: ensures that the victim will have a solvent entity that will indemnify, regardless of who is ultimately at fault.

Why vary from the fault-based liability system?

Should robots have rights and obligations?



https://www.youtube.com/watch?v=0VgxAnZKM14

Asimov's three laws of robotics:

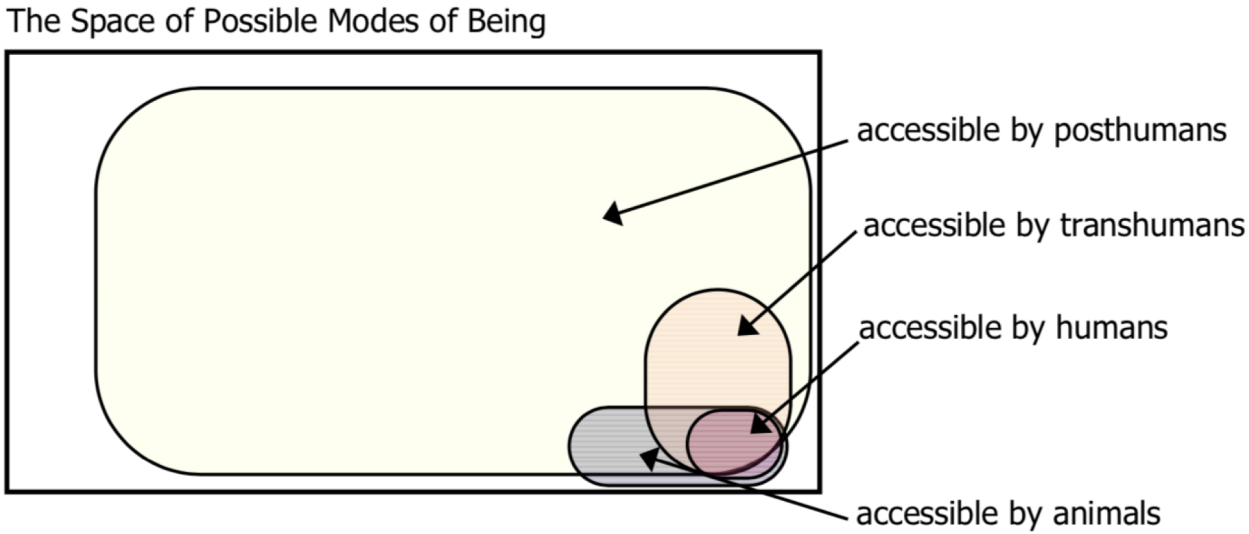
- **First Law**: A robot may not injure a human being or, through inaction, allow a human being to come to harm.
- **Second Law:** A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
- **Third Law:** A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

What about augmented humans?

Transhumanists view human nature as a **work-in-progress**, a half-baked beginning that we can learn to remold in desirable ways.

Current humanity need not be the endpoint of evolution.

Nick Bostrom, Transhumanist Ethics



Oviedo Convention on Human Rights and Biomedicine

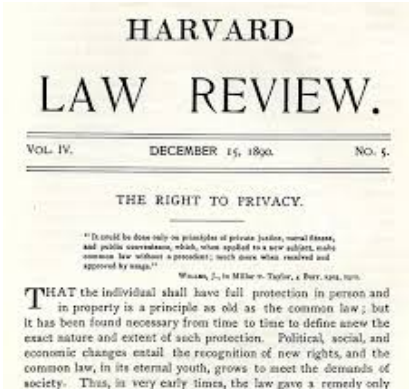
Article 13 – Interventions on the human genome
An intervention seeking to modify the human genome may only be undertaken for **preventive, diagnostic or therapeutic purposes** and only if its aim is not to introduce any modification in the genome of any descendants.

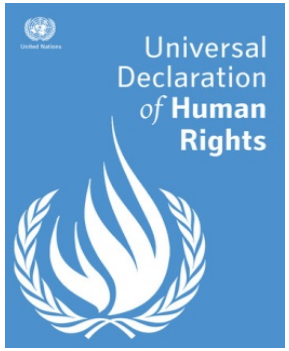
For legal and ethical resources, see the dedicated website of the Council of Europe

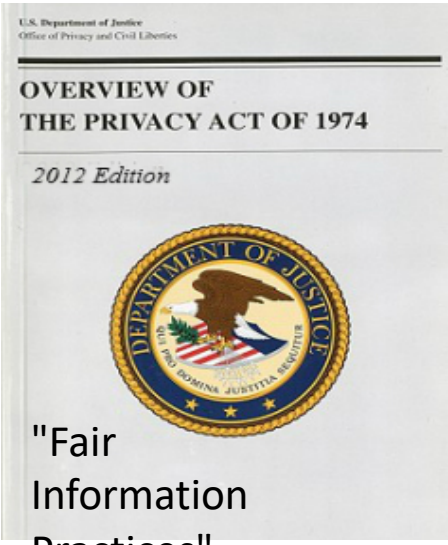
What's the difference between therapy and enhancement?

An introduction to data protection law

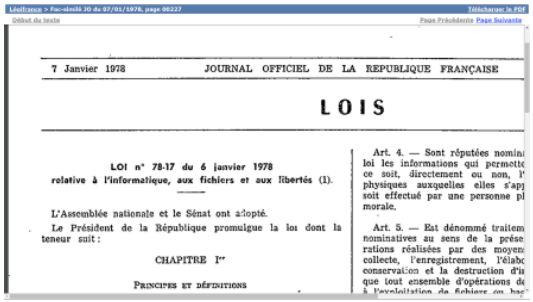
A brief history of privacy law





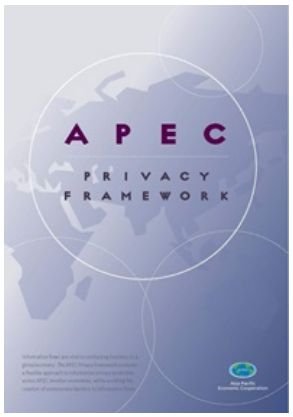


"Fair Information Practices"



OECD
BETTER POLICIES FOR BETTER LIVES

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE



CCPA 2018

1789: 4th Amendment of the US Constitution

1870: the Right to be left alone

1948: Universal Declaration of Human Rights

1950: The European Convention on Human Rights

The 1974 Privacy Act

1980: The OECD guidelines

1983: Council of Europe convention 108

1995: Directive 95/46

2000: EU Charter of Fundamental Rights

2016: GDPR **2018** CCPA

The GDPR co-exists with other privacy laws

The European General Data Protection Regulation 2016/679 (GDPR)

GDPR	"Police" Directive	E-Privacy Directive
• Applies to private companies and governments	• Applies to government processing for law enforcement	• Special rules for « cookies », and electronic communications

Charter of Fundamental Rights of the European Union
European Convention on Human Rights
Case law of the European Court of Justice and the European Court of Human Rights

What is personal data

A dynamic IP address: Directly or indirectly identifiable?	"Salted hashing" of data techniques. What does "directly or indirectly identify" mean?
• "*that would not be the case if the identification of the data subject was prohibited by law or practically impossible on account of the fact that it requires a disproportionate effort in terms of time, cost and man-power, so that the risk of identification appears in reality to be insignificant.*" **Breyer case (CJUE – 19 October 2016)**	• Salted hashing techniques "leave the data controller in a position to identify the data subjects and prohibit neither the correlation of records relating to the same individual nor the inferring of information concerning him" - **Identification = "singling out" - JC Decaux case (Conseil d'Etat- 8 février 2017)**

Special categories of data

- Explicit consent
- Obligations in the field of employment and social security/protection
- Legitimate and non-profitable activities
- Manifestly made public by the data subject
- Exercise or defence of legal claims
- Public interest, scientific or historical research purposes or statistical purposes
- Criminal convictions and offences

The GDPR architecture

One legal basis AND respect for all processing principles

- Legal basis (art. 6)

« or »

  - Consent
  - Necessary to perform a contract
  - Legal obligation
  - Vital interest
  - Public interest
  - Legitimate interest

- Respect of principles (art. 5)

  - Lawfulness, fairness and transparency
  - Purpose limitation
  - Data minimisation and accuracy
  - Storage limitation
  - Integrity and confidentiality
  - Accountability

« and »

Data subject rights

- More detailed provision of information
- Right of access
- Rights to rectification, erasure and restriction of processing
- Right to data portability
- Right to object to the processing
- Rights with respect to automated processing

Definition of ‘data controller’ and ‘data processor’

Data controller

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.**

Data processor

A natural or legal person, public authority, agency or other body which processes personal data **on behalf of the controller**.

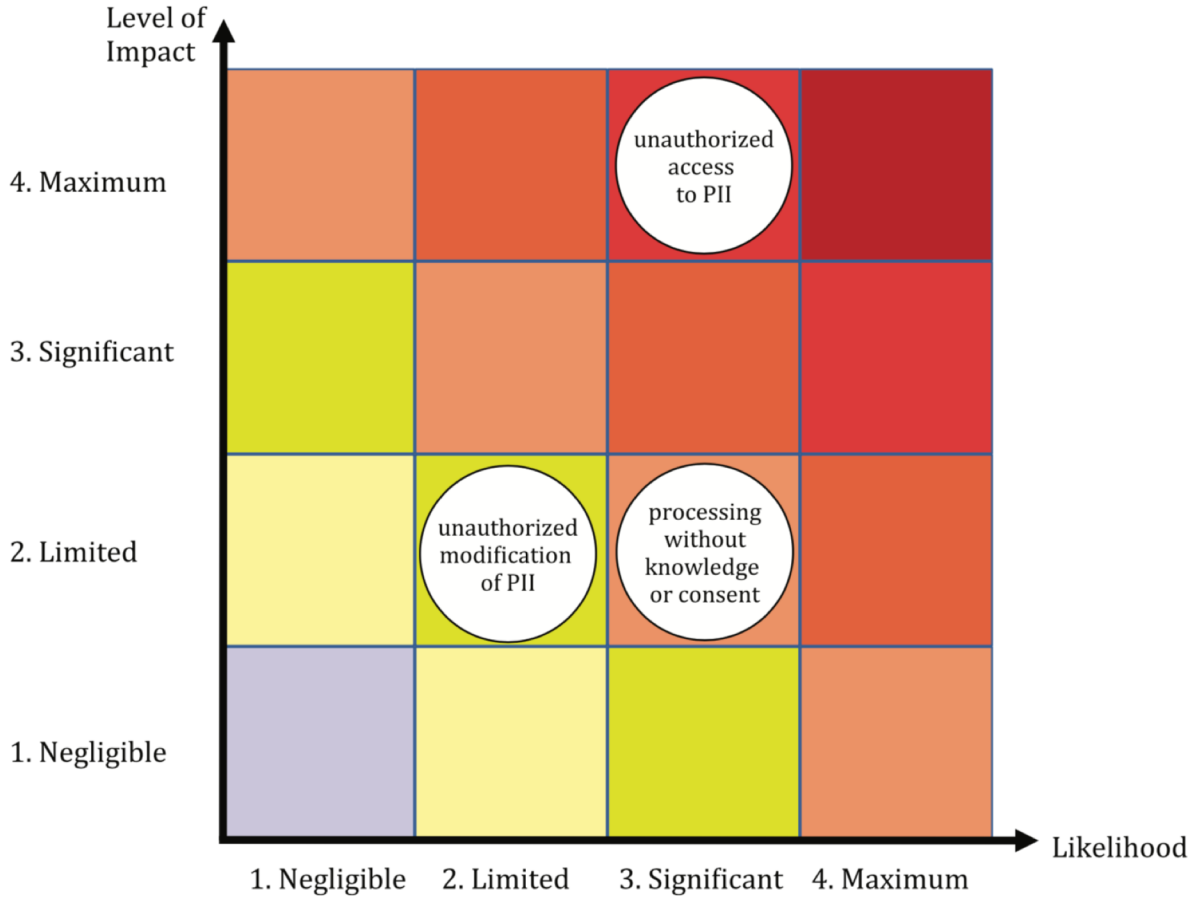
Other key principles of the GDPR

- Accountability
- Personal data register
- 'Appropriate measures'
- International transfers

Algorithmic decisions

- The right not to be subject to solely automated decisions
- The right to human intervention
- The right to receive *'meaningful **information about the logic involved**, as well as the significance and the envisaged consequences of such processing for the data subject'*

Privacy impact assessments



Source: ISO 29134

Figure D.2 — Example of a privacy risk map

International ‘soft law’ recommendations on AI ethics

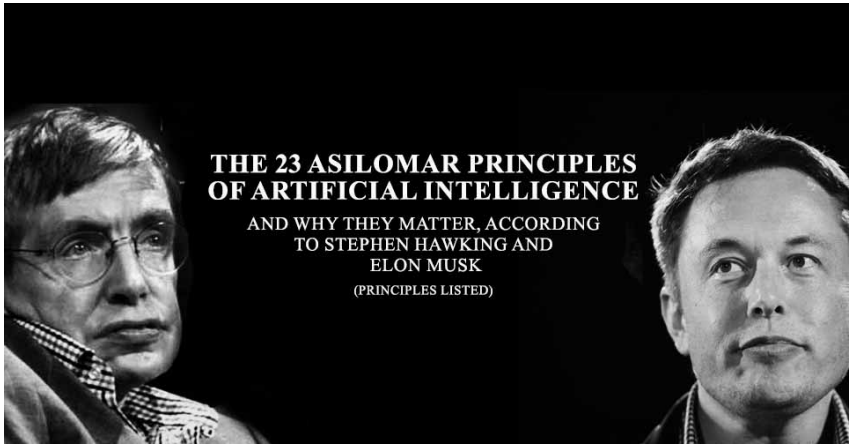
HLEG Guidelines

The Guidelines (2019) impose seven princip

- ✓Human agency and oversight
- ✓Technical Robustness and safety
- ✓Privacy and data governance
- ✓Transparency
- ✓Diversity, non-discrimination and fairness
- ✓Societal and environmental well-being
- ✓Accountability

Asilomar principles

2017 ‘Future of Life’ conference
23 principles



OECD and G20 recommendations

Five principles

- inclusive growth, sustainable development and well-being;
- human-centred values and fairness;
- transparency and explainability;
- robustness, security and safety;
- and accountability.

[OECD Recommendation of the Council on Artificial Intelligence]

AI ethics by design

Law and ethics mixed together

A European regulation on ethical AI?

Proposal of the European Parliament 20 october 2020

Proposal for a

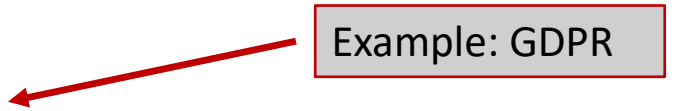
REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

on ethical principles for the development, deployment and use of artificial intelligence, robotics and related technologies

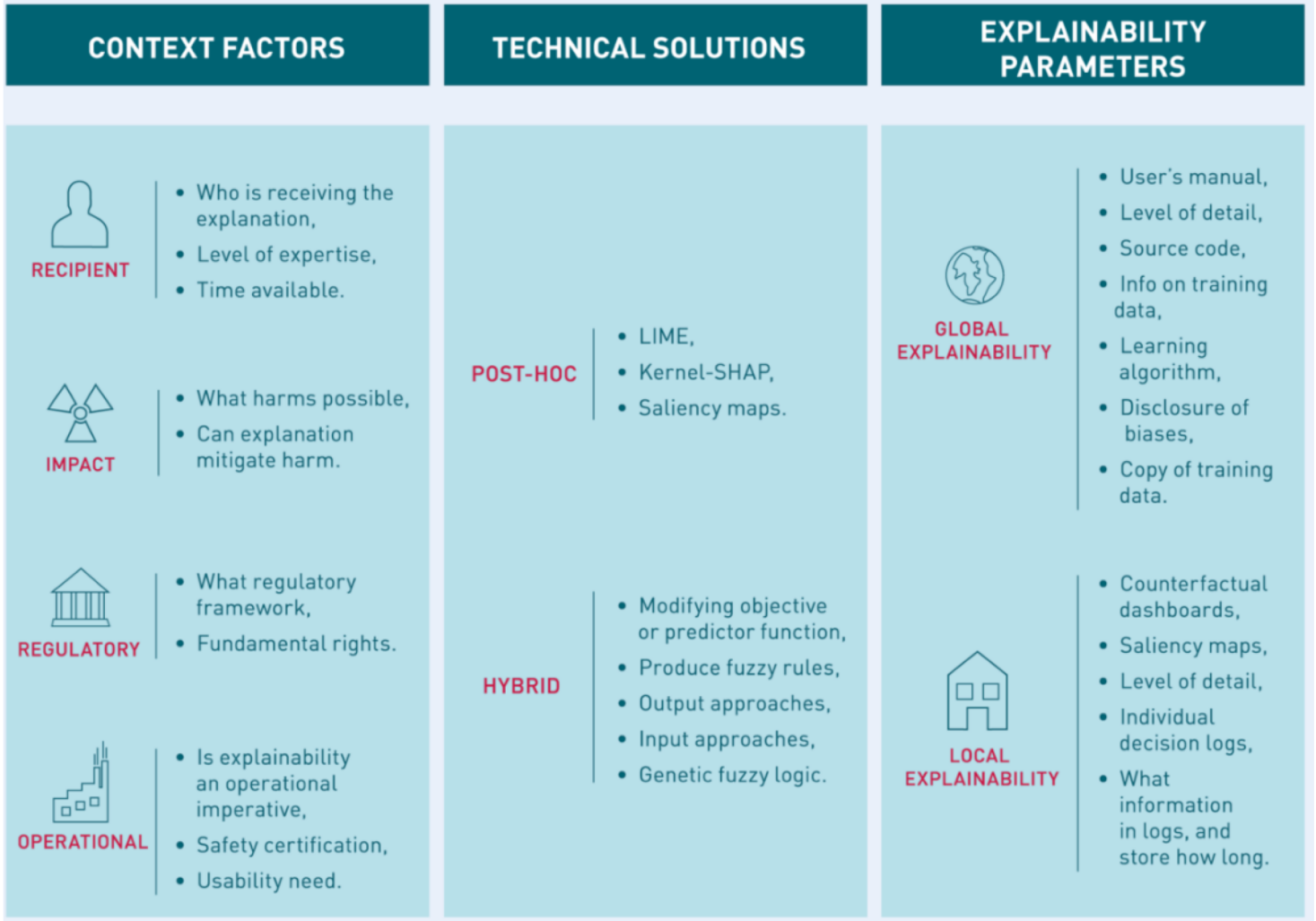
https://www.europarl.europa.eu/doceo/document/TA-9-2020-0275_EN.pdf

What are these principles trying to protect?

- Protect the individual from harm
  - Fundamental rights, safety, data protection, individual dignity and autonomy
- Protect society, democratic institutions, and human dignity in a collective sense
  - Possible cumulative effects of AI on the human ecosystem
    - Eg the destruction of democratic processes through *bots* and *fake news*
    - Autonomous lethal weapons
    - Transhumanism
- Manage innovation and risk uncertainty
  - Remove legal and risk uncertainties for responsible AI innovation



Explainability



Publications: Flexible and Context-Specific AI Explainability (arXiv:2003.07703)

Identifying the 'Right' Level of Explanation in a Given Situation (hal-02507316)

EXPLAINABILITY REQUIREMENTS

OPERATIONAL

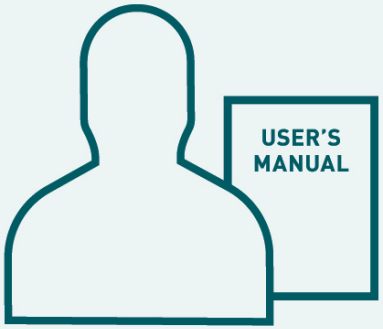
- Produce decisions with actionable insights,
- Safety certification,
- User trust,
- Making models more robust.

LEGAL AND REGULATORY

- The right to challenge decisions,
- Prevent discrimination,
- Protect privacy,
- Avoid systemic harms to society.

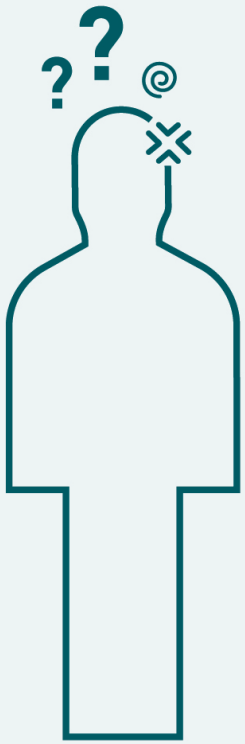
FROM TRANSPARENCY TO ACCOUNTABILITY

GLOBAL EXPLANATIONS



- Overview of the model,
- How it learned,
- What training data,
- Limitations to the model and use restrictions.

LOCAL EXPLANATIONS



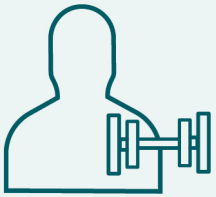
- Why was my application denied?
- What was the main factor?
- What can I change?
- Was I discriminated against?
- Why did you misclassify this image?

EXPLAINABILITY REQUIRES INFORMATION ABOUT THE LEARNING ALGORITHM AND THE TRAINED ALGORITHM

QUESTIONS TO ASK THE LEARNING ALGORITHM

- How did you learn?
- What's your objective function?
- What training and test data did you use?
- How did you tune the model?

QUESTIONS TO ASK THE TRAINED ALGORITHM



- What were the main factors leading to your decision?
- How would the decision change if we altered one of the factors?
- Show me a map of how you reasoned in this case.

TECHNICAL APPROACHES

POST-HOC APPROACHES

- Input perturbation,
- Saliency maps.

Explanor model produces local explanations by approximating the black box function.

Black box model remains black, and continues to make predictions without explanation.

PUTTING EXPLAINABILITY IN THE MODEL

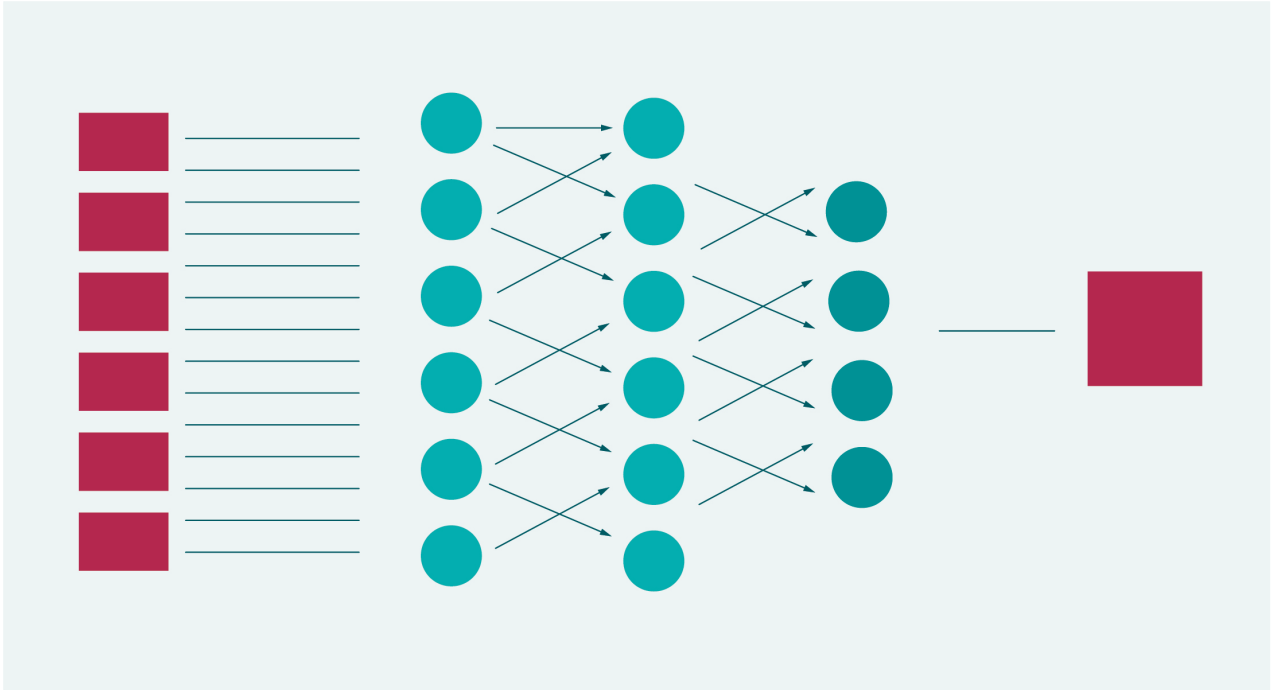
Modifying the black box itself to make it more understandable.

THREE APPROACHES TO EXPLAINABILITY BY DESIGN

1
Pre-process inputs based on knowledge rules.

2
Reward the network for following interpretable rules and representations.

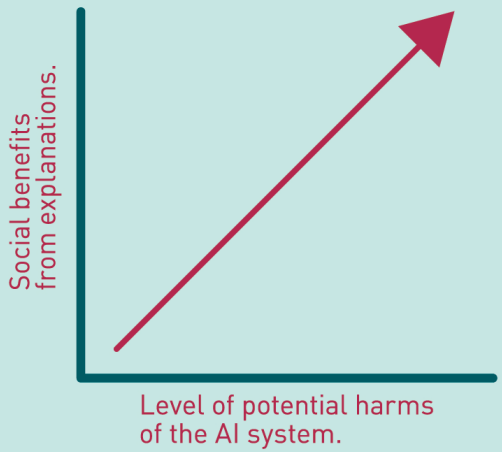
3
Filter outputs to exclude those that fall outside approved scenarios.



BENEFITS AND COSTS OF EXPLANATIONS

BENEFITS

Benefits of explanation generally increase with the level of potential harm caused by AI.



COSTS

- Design and operations costs,
- Reverting to less-performing models,
- Creating and storing logs,
- Interference with trade secrets,
- Interference with other rights, e.g. data protection.

EXAMPLE OF A HYBRID APPROACH

Knowledge rules tell the network what part of the image to focus on when learning.

X-RAY OF A KNEE	NEURAL NETWORK	PREDICTION