Justice, Fairness and Artificial Intelligence

Jean-Gabriel Ganascia

19 November 2021

Sorbonne University, LIP6 (computer science lab)

Ex-Chairman of the COMETS (CNRS Ethics Committee)

Jean-Gabriel.Ganascia@lip6.fr

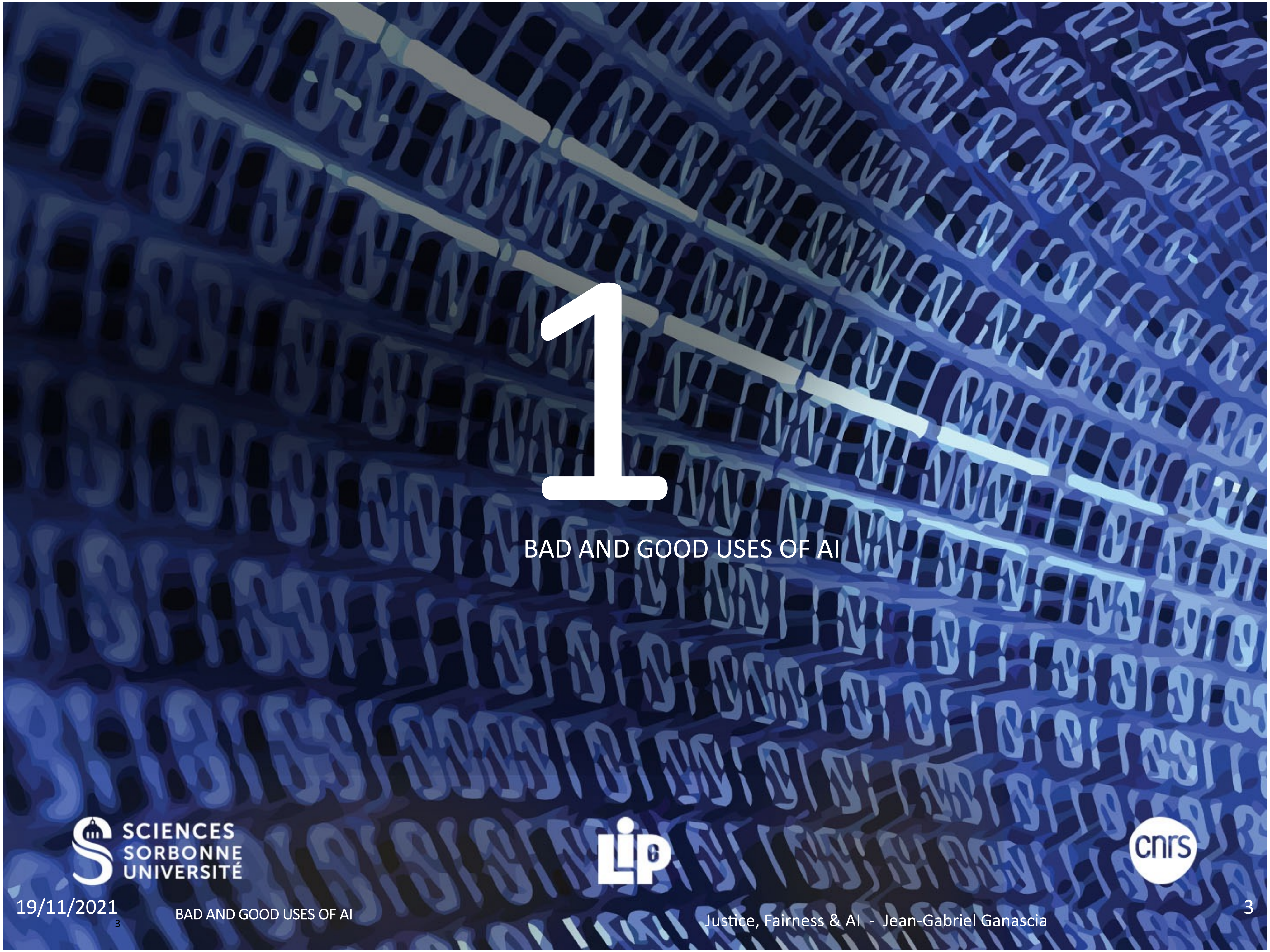
SCIENCES
SORBONNE
UNIVERSITÉ

LIP6

cnrs

Synoptic

1. Bad and Good Uses of AI
2. Establishing norms and regulations in AI
3. Ethics, Norms, Laws and Regulation
4. Justice and Fairness
5. Computational Ethics: Legal and Ethical supervisor



1

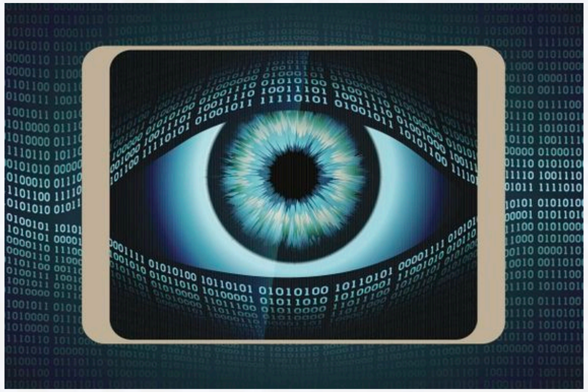
BAD AND GOOD USES OF AI

Misuses of AI: examples

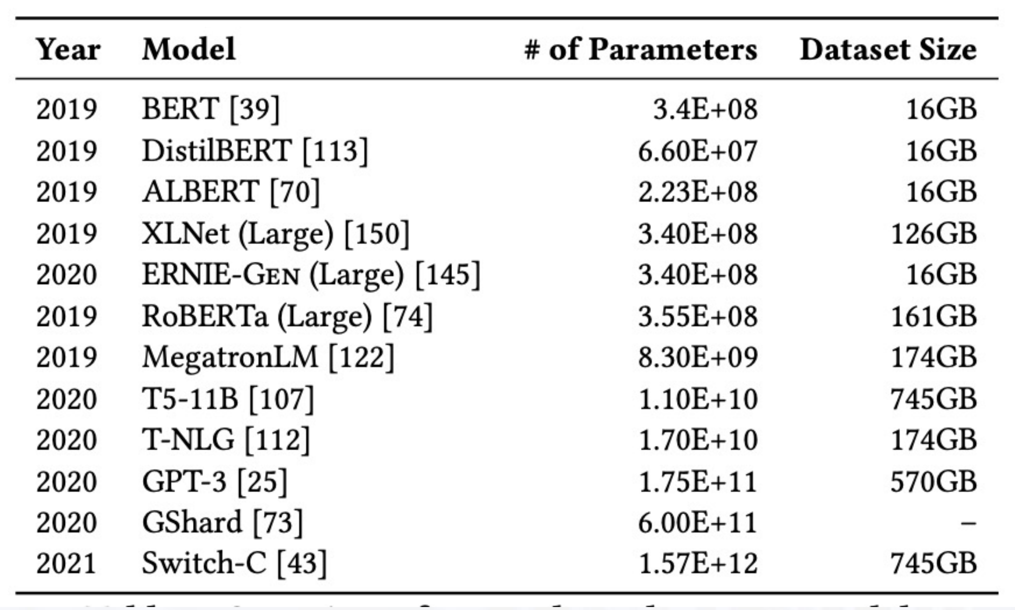
Irresponsible, Unjust and Unfair uses of AI

Use of AI that could infringe human dignity and autonomy

- Surveillance systems that would track every move — social credit in China
- Biased AI systems that are discriminatory — facial recognition
- Cast public opprobrium on those who disobey the rules
- AI text generation engines or image synthesis that could produce fake
- AI-based targeting dissemination techniques of these fake news.
- …

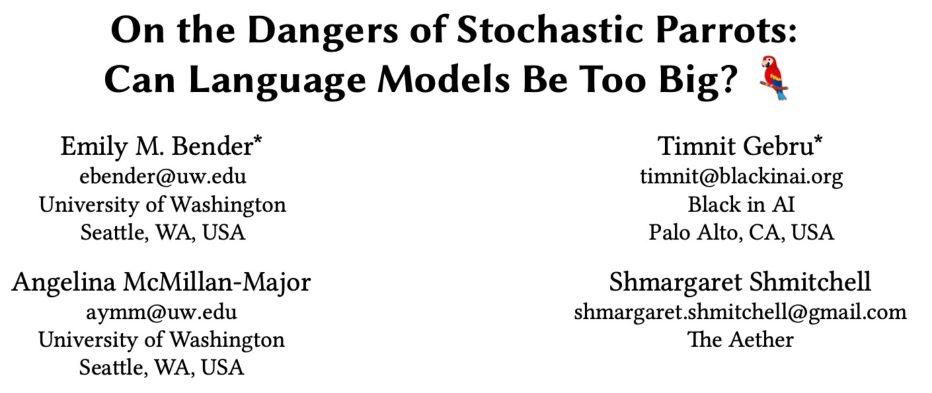


On the Dangers of Stochastic Parrots



Year	Model	# of Parameters	Dataset Size
2019	BERT [39]	3.4E+08	16GB
2019	DistilBERT [113]	6.60E+07	16GB
2019	ALBERT [70]	2.23E+08	16GB
2019	XLNet (Large) [150]	3.40E+08	126GB
2020	ERNIE-Gen (Large) [145]	3.40E+08	16GB
2019	RoBERTa (Large) [74]	3.55E+08	161GB
2019	MegatronLM [122]	8.30E+09	174GB
2020	T5-11B [107]	1.10E+10	745GB
2020	T-NLG [112]	1.70E+10	174GB
2020	GPT-3 [25]	1.75E+11	570GB
2020	GShard [73]	6.00E+11	–
2021	Switch-C [43]	1.57E+12	745GB

- Big Language Models
- Trained With Massive Texts (*encyclopedia*)
- Neural Networks with Trillions Parameters



On the Dangers of Stochastic Parrots:
Can Language Models Be Too Big? 🦜

Emily M. Bender*
ebender@uw.edu
University of Washington
Seattle, WA, USA

Timnit Gebru*
timnit@blackinai.org
Black in AI
Palo Alto, CA, USA

Angelina McMillan-Major
aymm@uw.edu
University of Washington
Seattle, WA, USA

Shmargaret Shmitchell
shmargaret.shmitchell@gmail.com
The Aether

On the Dangers of Stochastic Parrots

- Huge financial costs of language models
- Disastrous energy balance of learning!
- Learning biases
  - Corpus used (online collaborative encyclopedias): reflection of the “white male” dominant thought which does not reflect minorities
- Filtering necessary to avoid abuses (like Microsoft Tay's):
  - at the same time that it eliminates pornography and incitement to hatred, it eliminates LGBT sites...

Examples of useful Applications: Medical Aspects

Processing Huge Masses of Medical Data

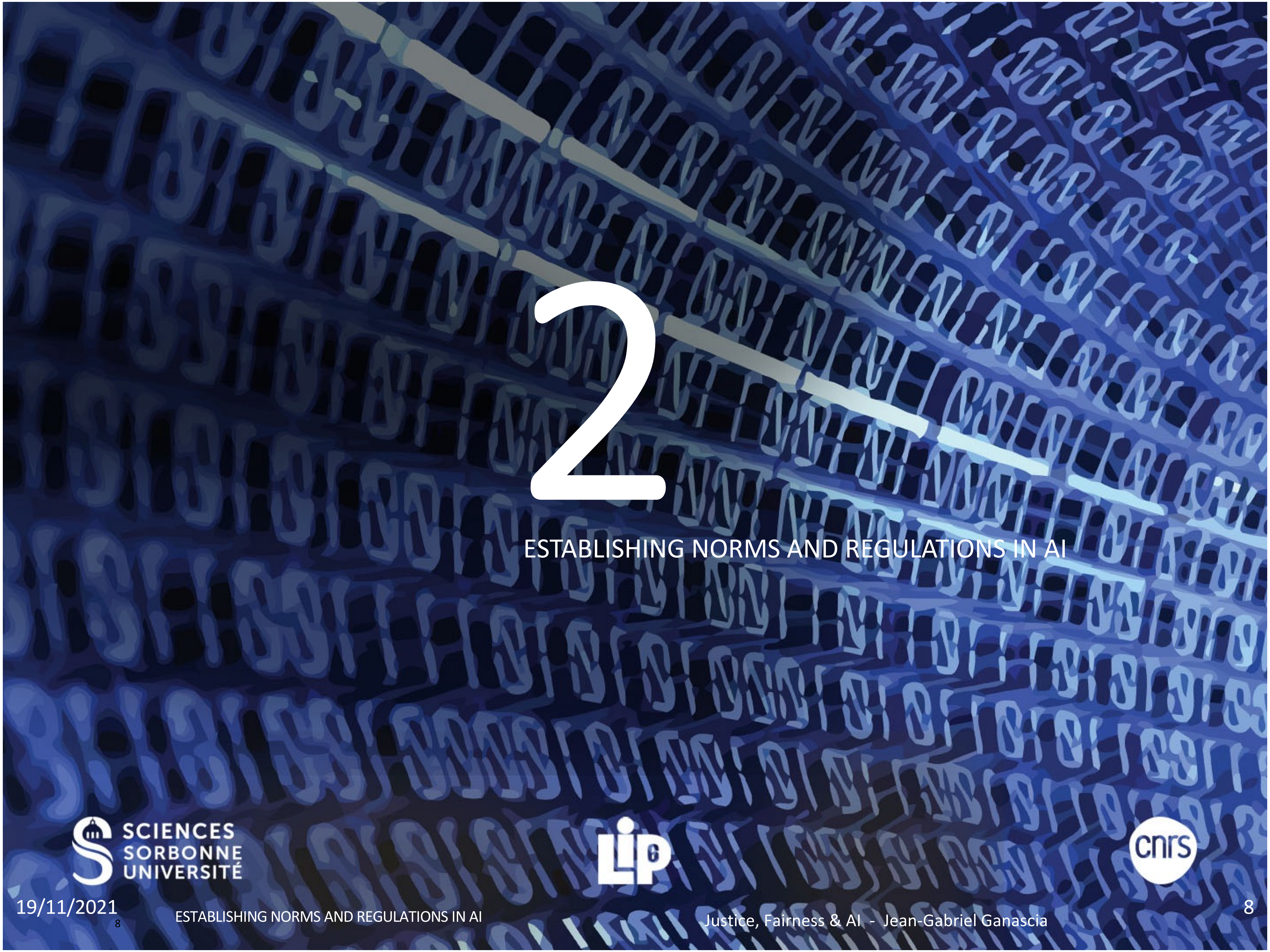
- Extracting medical knowledge from patient data (X-rays, clinical signs, etc.)
- Extracting biological information (e.g. genetic factors explaining the evolution of the disease, etc.)

Bioinformatics

- Modeling biological processes (e.g. mechanism of introduction of the virus into cells, genetic factors explaining the evolution of the disease, etc.)

Extraction of Knowledge from the Scientific Literature

- More than 60,000 papers on CoViD-19 were produced last 6 months!



2

ESTABLISHING NORMS AND REGULATIONS IN AI

GDPR- *General Data Protection Regulation*

Principles

- **Finality**: an organization must present a legitimate objective for collecting personal data
- **Transparency**: an organization must notify users about the collection and sharing of information with third parties
- **Respect of Personal Rights**: the user has the right to accept or reject data collection. They can also ask for their data to be corrected and permanently deleted

Regulation of Artificial Intelligence

The European Commission's

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

DRAFT

ETHICS GUIDELINES FOR TRUSTWORTHY AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

1. Autonomy
2. Beneficence
3. Non-Maleficence
4. Justice

+

5. Transparency

ems (AIS) must permit the growth of the

PRINCIPLE

ople's autonomy, and with the goal of
urroundings

ND INTIMACY PRINCIPLE

Concepts and Principles Invoked

- Justice
- Fairness
- Lawfulness
- Transparency
- Non-discrimination
- Human Autonomy
- Prevention of Harms
- Human Agency
- Respect Privacy
- …

The European Commission's

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

AI

DRAFT

ETHICS GUIDELINES FOR TRUSTWORTHY AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

Origin of these Concepts and Principles

- Fundamental Rights

UN Universal Declaration of Human Rights (1948)

Right to

- self-determination
- liberty
- due process of law
- freedom of movement
- privacy
- freedom of though
- freedom of religion
- freedom of expression
- peaceful assembly
- freedom of association

- Belmont Report (1978)

Ethical Principles and Guidelines for the Protection of Human Subjects of Research

*The National Commission for the **Protection of Human Subjects** of **Biomedical** and **Behavioral Research***

- Autonomy
- Beneficence
- Non Maleficence
- Justice

Based on Human Rights & Bioethics Principles

Trustworthy AI

1. Lawful
2. Ethical
3. Robust

Three layers

1. Principles:
   - Respect for Human Autonomy
   - Prevention of Harms
   - Fairness
   - Explicability
2. Realizing Trustworthy AI
   - **Seven Requirements**: human agency, technical robustness, privacy, transparency, non-discrimination and fairness, societal and environmental well-being, accountability
   - **Technical and non technical methods**
3. Assessing Trustworthy AI

The European Commission's

High-Level Expert Group on Artificial Intelligence

DRAFT

Ethics Guidelines for Trustworthy AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

3

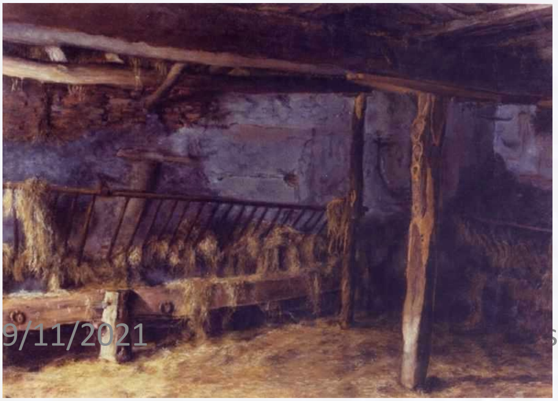
ETHICS, NORMS, LAWS AND REGULATION

Moral and Ethics

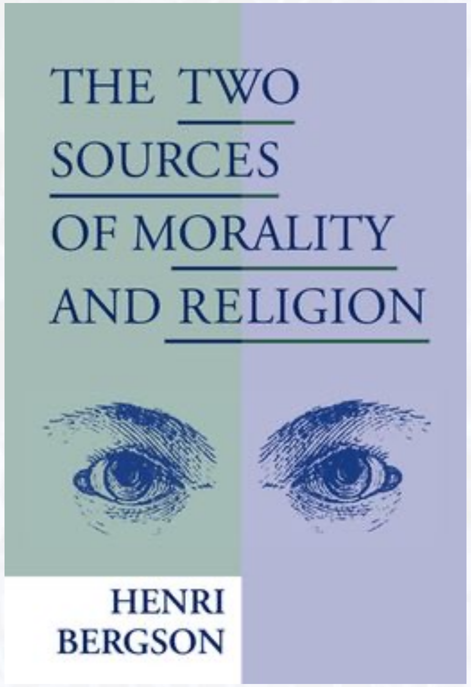
Ethics: Latin ethica; Greek *êthikos*, *êthikê*, from *êthos*, ‘custom’, ‘mores’

Originally, in Greek, *êthos* meant a place familiar to animals, e.g. a stable.

With Aristotle, means the rational deliberation necessary to act well.



Moral : Latin *moralis* from *mores* → Mores



THE TWO SOURCES OF MORALITY AND RELIGION

HENRI BERGSON

The art (or the science?) of directing one's conduct

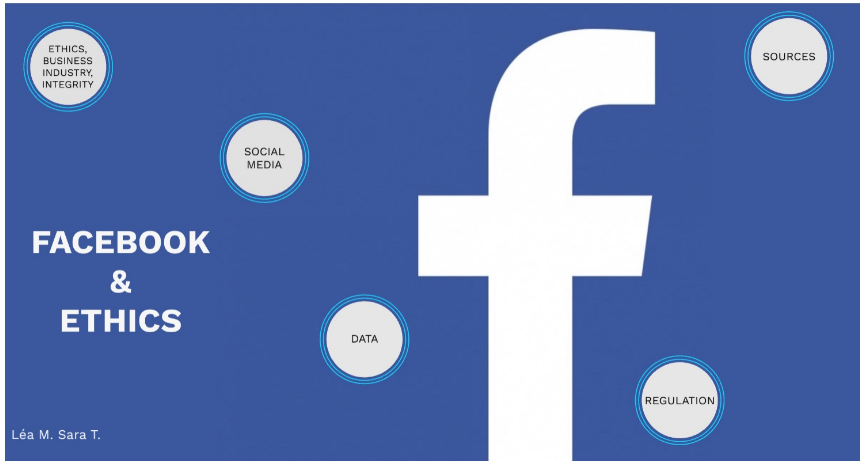
Do not confuse Ethics with Laws, Regulation and Norms

- **Laws**
  - **Right**: set of human laws
    - Distinction between human and natural laws
  - **Laws are voted** (Parliaments)
  - **Authority of the law**: sanction
  - **Law enforcement**: what is allowed and what is not

- **Regulation**
  - **Administrative rules** that clarify laws
- **Norms**:
  - **Mandatory rules** that do not necessarily come from the law (e.g. industrial standards, environmental rules)

Norms, Politics and Power

- Presence of GAFAMI in standardization institutions

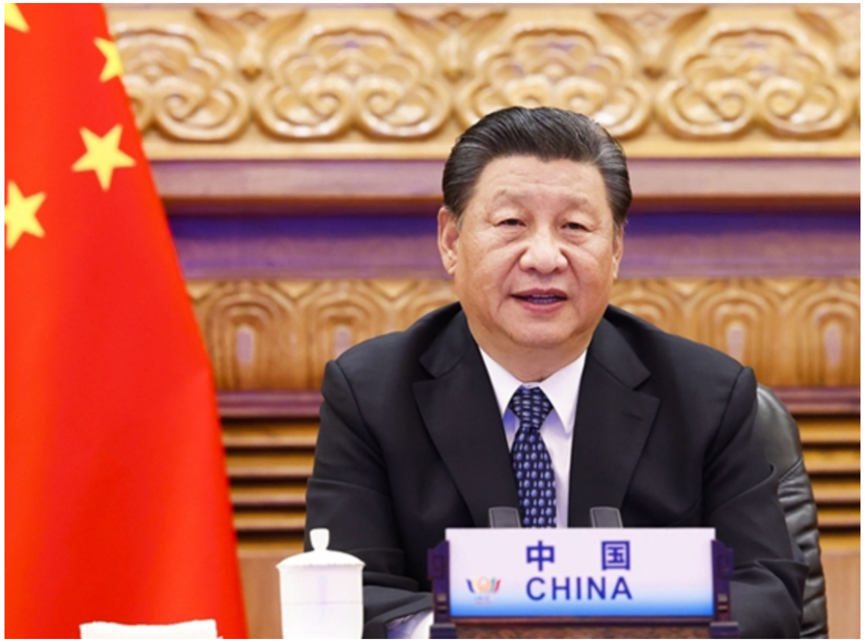


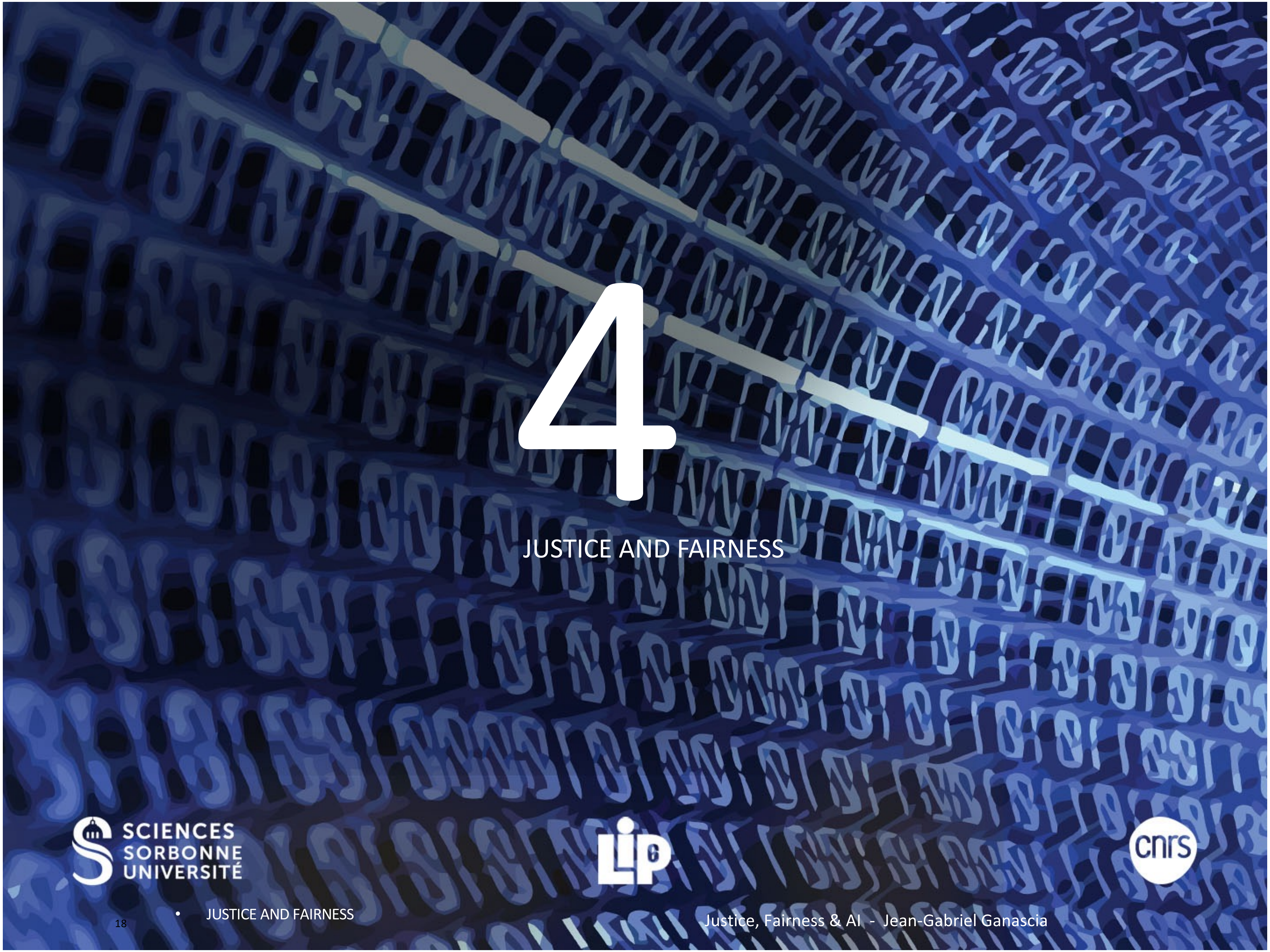
- Appearance of China

XI Jinping

5G first evolutionary standard announced **completion of Chinese wisdom into international standards**

(5G首个演进标准宣布完成 中国智慧融入国际标准), People's Daily Online, Author: Zhao Chao *(人民网), 4 July 2020*





4

JUSTICE AND FAIRNESS

Concepts and Principles

- Justice
- Fairness
- Lawfulness
- Transparency
- Non-discrimination
- Human Autonomy
- Prevention of Harms
- Human Agency
- Respect Privacy
- …

The European Commission's

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

DRAFT

ETHICS GUIDELINES FOR TRUSTWORTHY AI

Working Document for stakeholders' consultation

Brussels, 18 December 2018

Focus on Justice and Fairness

Justice

- Lawfulness
- Rights
- Human Rights

Fairness

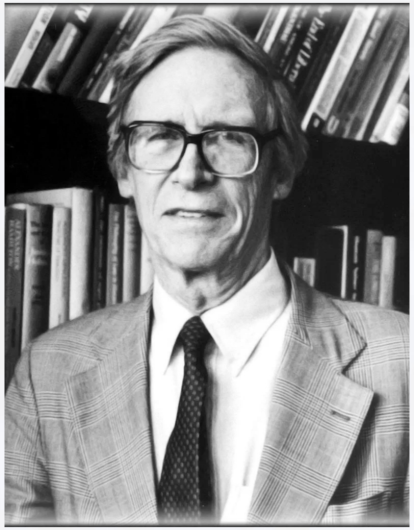
- Non discrimination
- Impartiality

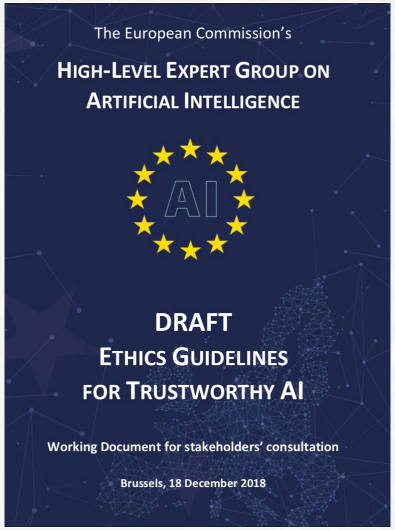
Fairness Principle (HLEG AI)

- ensure **equal** and **just** distribution of both benefits and costs,
- ensure that individuals and groups are **free from bias**, **discrimination** and **stigmatisation**

Principle of Justice: “Be Fair”

*Developers and implementers need to ensure that individuals and minority groups maintain freedom from **bias**, stigmatization and **discrimination***





The European Commission’s

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE

AI

DRAFT

ETHICS GUIDELINES FOR TRUSTWORTHY AI

Working Document for stakeholders’ consultation

Brussels, 18 December 2018

Justice as Fairness – John Rawls

Equal distribution of opportunities

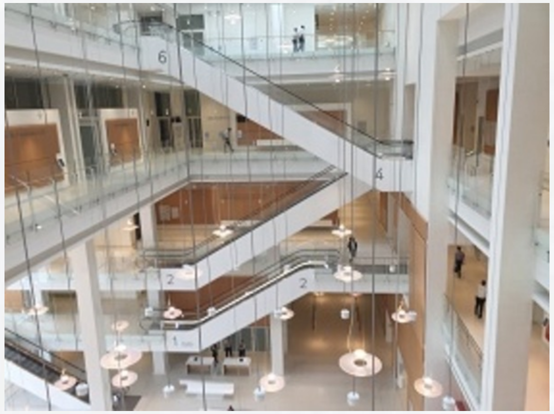
- ***bias**: prejudice for or against something or somebody, that may result in unfair decisions.*
- ***discrimination**: concerns the variability of AI results between individuals or groups of people based on the exploitation of differences in their characteristics that can be considered either intentionally or unintentionally (such as ethnicity, gender, sexual orientation or age), which may negatively impact such individuals or groups*

Lady Justice Wears a Blindfold

- Allegory of the Impartiality of Justice
- Are Data and Algorithms impartial?
- Are Machine Free of Dogmas and Bias?

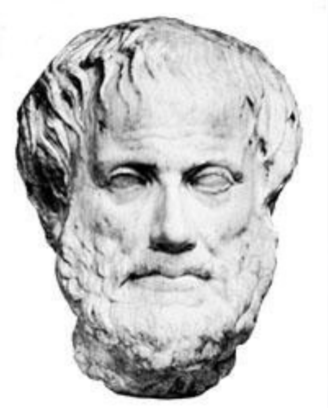
Justice & Just — Equity & Equality

- **Justice**:
  - Institution: judges, etc.
  - Set of laws
  - Justice applies the laws equally to everybody



- **Just**:
  - Correction to the law
  - “between the legal and the good” Paul Ricœur

- **Equality**
  - Give the same to everybody → distributivity
- **Equity**
  - Distribute according the need
  - The equity corrects the Law

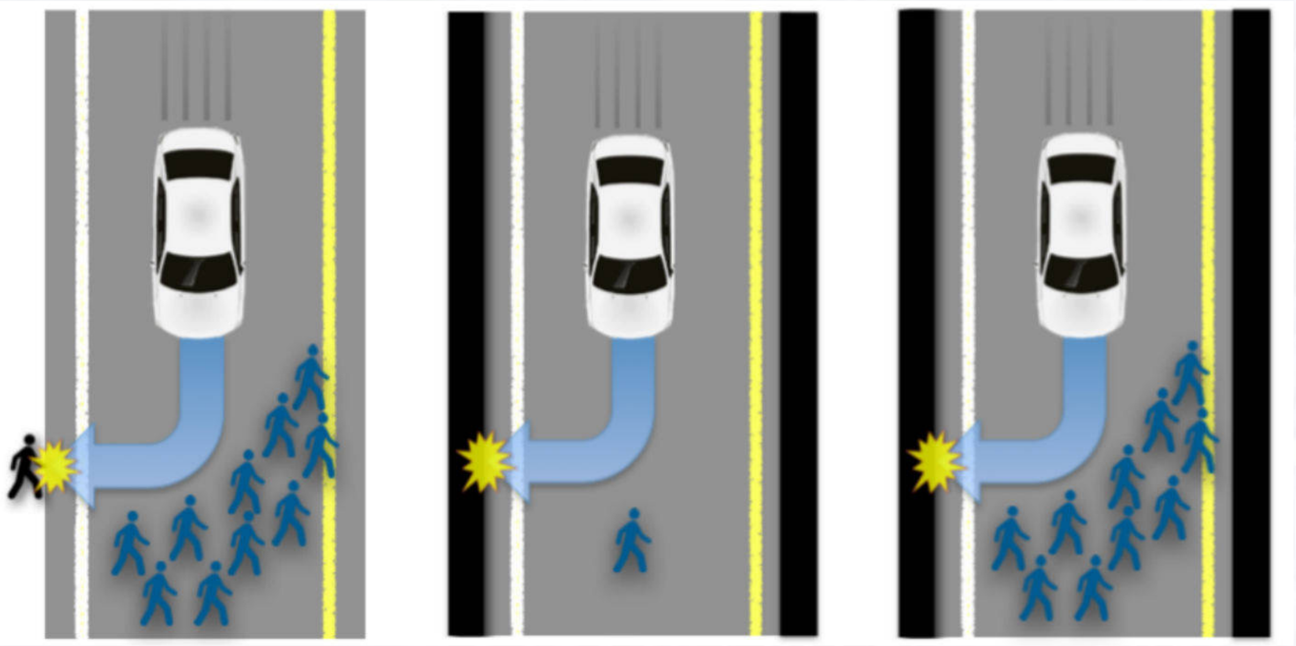


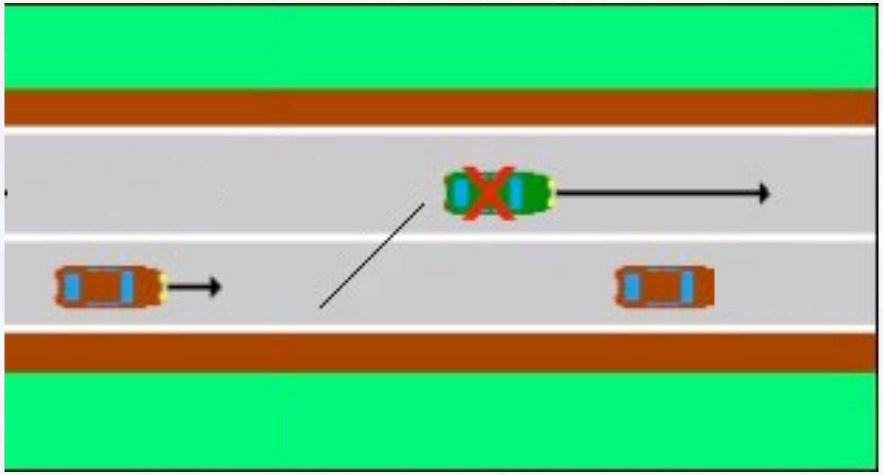
5

COMPUTATIONAL ETHICS

LEGAL AND ETHICAL SUPERVISOR

Deliberation: programming an ethical supervisor





Domain: data manipulation - GDPR

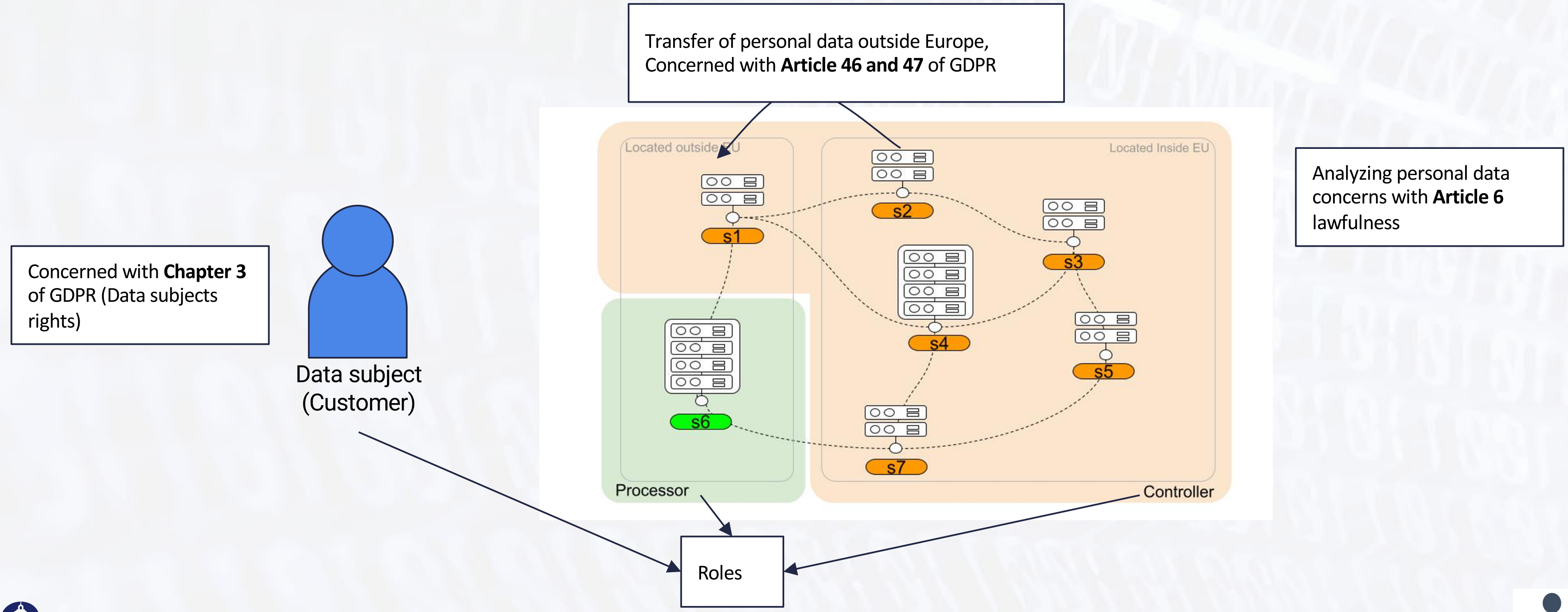
An **international European company** operates in multiple **EU** countries and as well as **US**.

Each sector owns a **server** only for **storing personal data** e.g. s1, s3,...

The servers are connected through an internal network and can transmit data among each other.

Two of the servers are **data processors** (**S4, S6**) in which the company analyses customers data.

A customer is **data subject** who has given her consent for a series of processing



Transfer of personal data outside Europe, Concerned with **Article 46 and 47** of GDPR

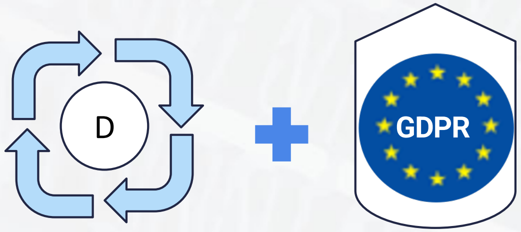
Analyzing personal data concerns with **Article 6** lawfulness

Concerned with **Chapter 3** of GDPR (Data subjects rights)

Data subject (Customer)

Roles

Data Manipulation Planning & Legal and Ethical Compliance Checking



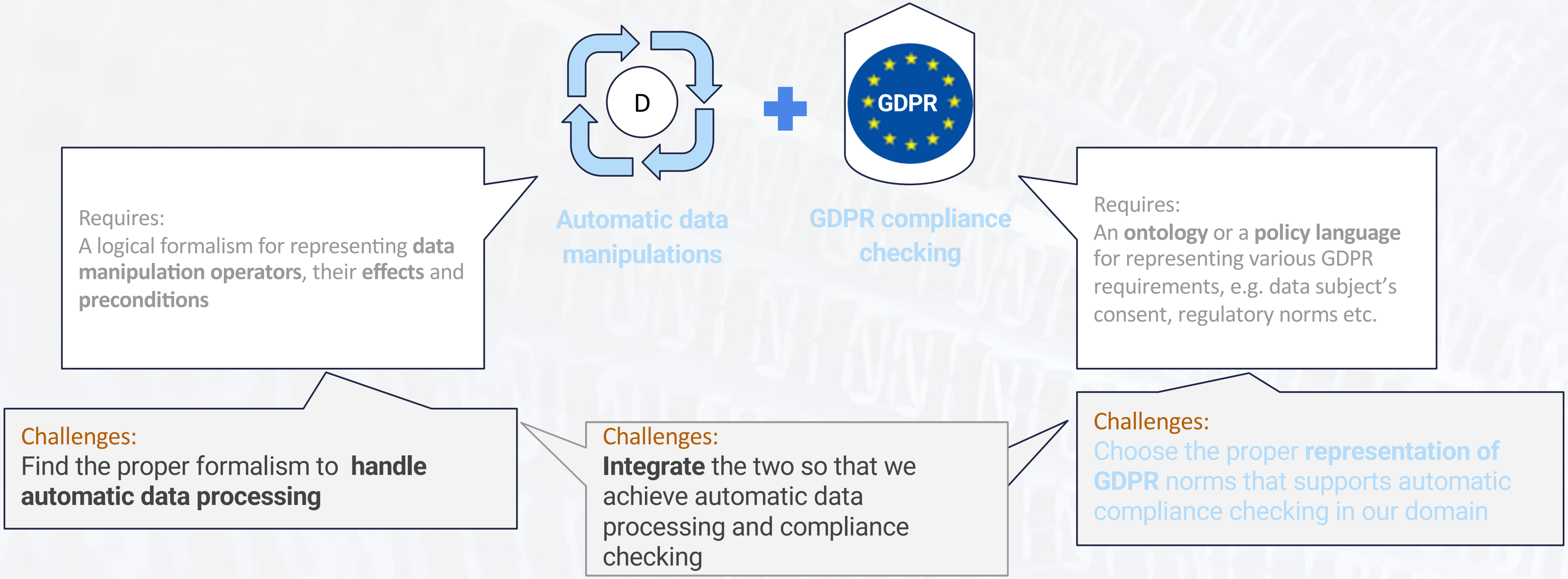
Automatic data manipulation

GDPR compliance checking

Requires:
A logical formalism for representing **data manipulation operators**, their **effects** and **preconditions**

Requires:
An **ontology** or a **policy language** for representing various GDPR requirements, e.g. data subject's consent, regulatory norms etc.

Data Manipulation Planning & Legal and Ethical Compliance Checking



Requires:
A logical formalism for representing **data manipulation operators**, their **effects** and **preconditions**

Automatic data manipulations

GDPR compliance checking

Requires:
An **ontology** or a **policy language** for representing various GDPR requirements, e.g. data subject's consent, regulatory norms etc.

Challenges:
Find the proper formalism to **handle automatic data processing**

Challenges:
Integrate the two so that we achieve automatic data processing and compliance checking

Challenges:
Choose the proper **representation of GDPR** norms that supports automatic compliance checking in our domain

Research on Realtime Compliance Mechanism for AI (RECOMP)

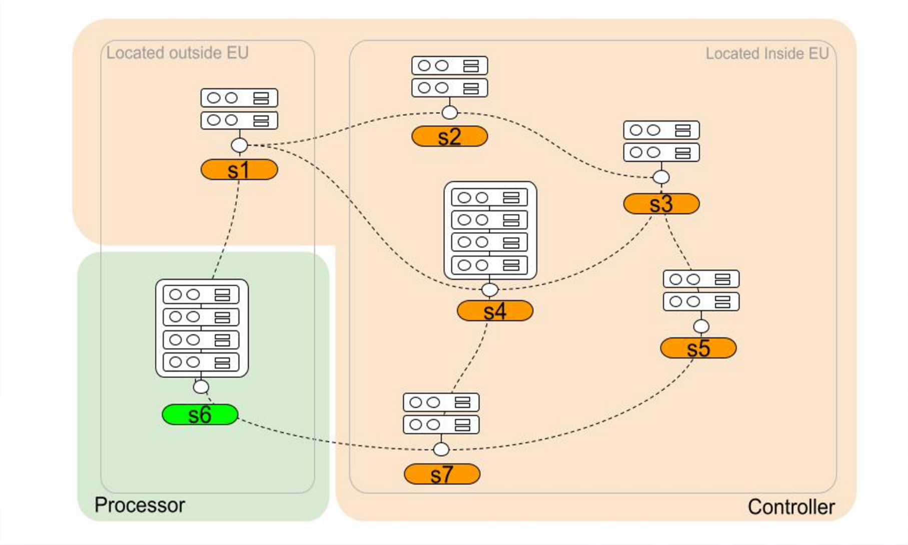
an International Project (France – Germany - Japan)

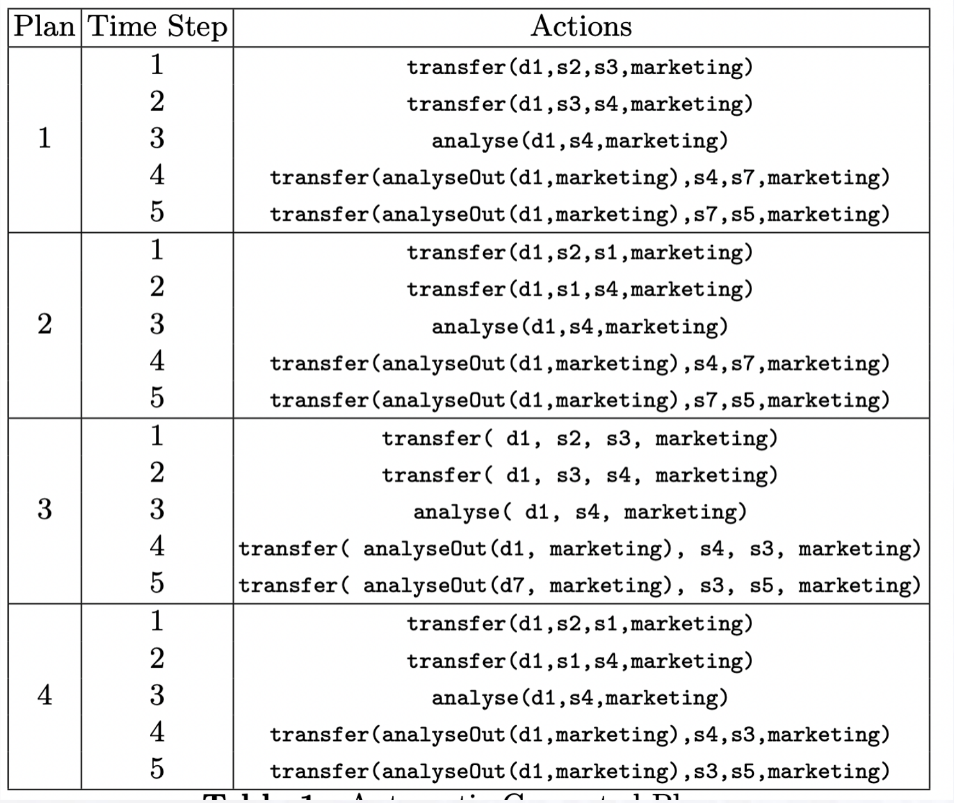
Evaluation

Initial state: d1 is located as storage s2

Goal state: output of the analysis should be at storage s5

Generated Plans (the table)

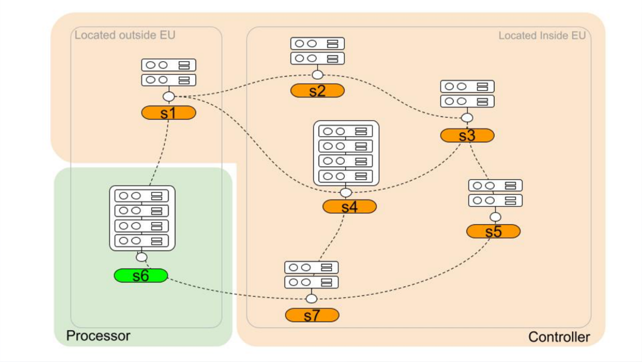


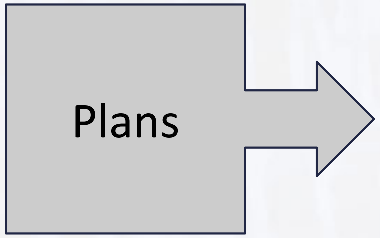


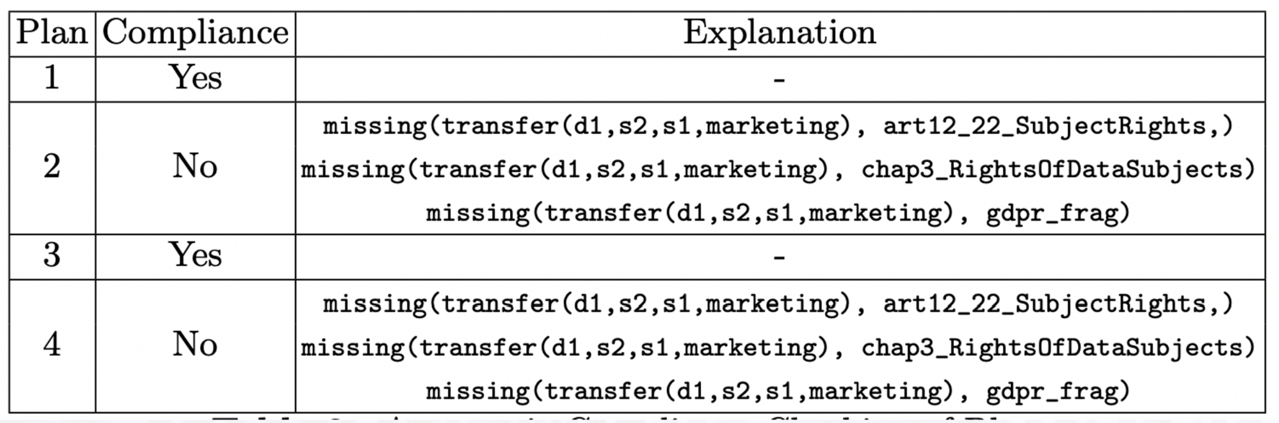
Plan	Time Step	Actions
1	1	transfer(d1,s2,s3,marketing)
	2	transfer(d1,s3,s4,marketing)
	3	analyse(d1,s4,marketing)
	4	transfer(analyseOut(d1,marketing),s4,s7,marketing)
	5	transfer(analyseOut(d1,marketing),s7,s5,marketing)
2	1	transfer(d1,s2,s1,marketing)
	2	transfer(d1,s1,s4,marketing)
	3	analyse(d1,s4,marketing)
	4	transfer(analyseOut(d1,marketing),s4,s7,marketing)
	5	transfer(analyseOut(d1,marketing),s7,s5,marketing)
3	1	transfer(d1, s2, s3, marketing)
	2	transfer(d1, s3, s4, marketing)
	3	analyse(d1, s4, marketing)
	4	transfer(analyseOut(d1, marketing), s4, s3, marketing)
	5	transfer(analyseOut(d7, marketing), s3, s5, marketing)
4	1	transfer(d1,s2,s1,marketing)
	2	transfer(d1,s1,s4,marketing)
	3	analyse(d1,s4,marketing)
	4	transfer(analyseOut(d1,marketing),s4,s3,marketing)
	5	transfer(analyseOut(d1,marketing),s3,s5,marketing)

Evaluation

Compliance results + Explanation







Plan	Compliance	Explanation
1	Yes	-

2	No	missing(transfer(d1,s2,s1,marketing), art12_22_SubjectRights,)
		missing(transfer(d1,s2,s1,marketing), chap3_RightsOfDataSubjects)
		missing(transfer(d1,s2,s1,marketing), gdpr_frag)

3	Yes	-

4	No	missing(transfer(d1,s2,s1,marketing), art12_22_SubjectRights,)
		missing(transfer(d1,s2,s1,marketing), chap3_RightsOfDataSubjects)
		missing(transfer(d1,s2,s1,marketing), gdpr_frag)

“Ethical” Artificial Agent

Classical Kantian distinction between

- Acting *from duty*

and

- Acting *in accordance with duty*

- “Ethical” Artificial Agents are only acting *in accordance with duty*, because they have no proper motivation

Thank You!

SCIENCES
SORBONNE
UNIVERSITÉ

LIP6

cnrs