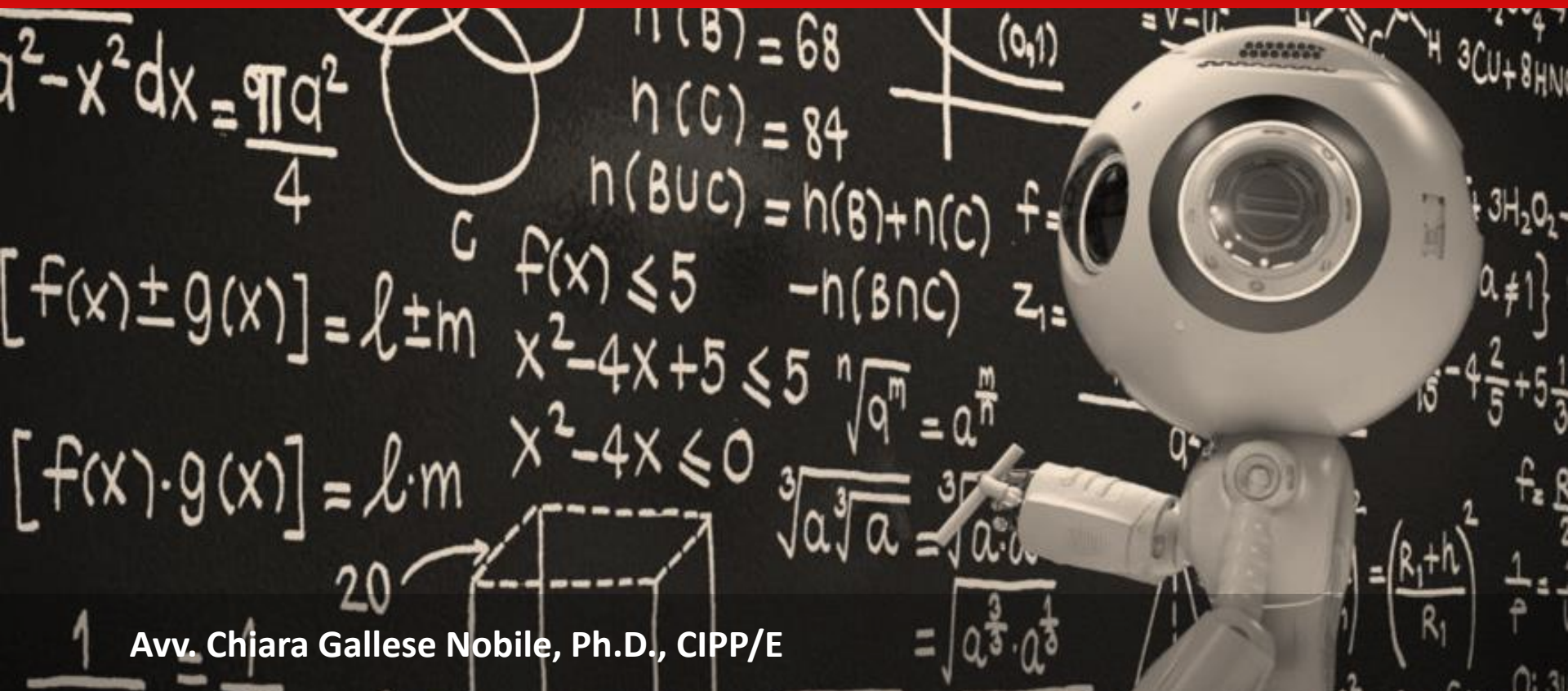


# Ethical and legal aspects of Data Science



**Avv. Chiara Gallese Nobile, Ph.D., CIPP/E**

# Project 101108151 — DataCom — HORIZON-MSCA-2022-PF-01

## A new EU Framework for an Ethical Re-use of Health Data



Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them.



**Funded by  
the European Union**

# The project



DataCom has started on the 01/07/2023 at University of Turin (Turin, Italy) at the Department of Law, under the supervision of Prof. Ugo Pagallo.

It aims to build a new framework to facilitate an ethical secondary use of health data held by public bodies to improve accountability and enhance responsible re-use.

To build it, I will develop and test in intersectoral practical environments the innovative concept of “Ethical Commodification”: the possibility of exploiting personal data in an ethical way for the public good, in accordance with the data subject’s expectations and needs and taking into account the risks associated with the exploitation of anonymized data sets.

I will focus on 3 Member States: Italy, Spain, and the Netherlands.

# Outline

Basic principles of AI Auditing

Basic principles of AI Ethics

The new AI Act proposal

Applying ethical principles to data sets: practical examples

# I. BASIC PRINCIPLES OF AI AUDITING



# What is AI auditing?

AI Auditing is the **review** of the whole AI lifecycle, including the post-market phase, with the aim of identifying and addressing **risks** for citizens and the society. It is carried out through applying legal, ethical and technical principles in practice.

*«An ethics-based auditing can improve the quality of decision making, increase user satisfaction, unlock growth potential, enable law-making, and relieve human suffering»*

- Mokander & Floridi

# How to perform AI auditing?

An AI Audit is carried out through the practical implementation of several principles:

- **Ethics:** checking if the consequences on individuals and society have been taken into account
- **Human Rights:** checking if fundamental rights of citizens are respected
- **Law:** checking if legal compliance has been performed (privacy, cybersecurity, national laws, etc.)
- **Technical good practices:** checking if the system is sound, safe, secure, correct, and biasless (e.g.: avoid sample bias)

# Guidance to perform AI auditing

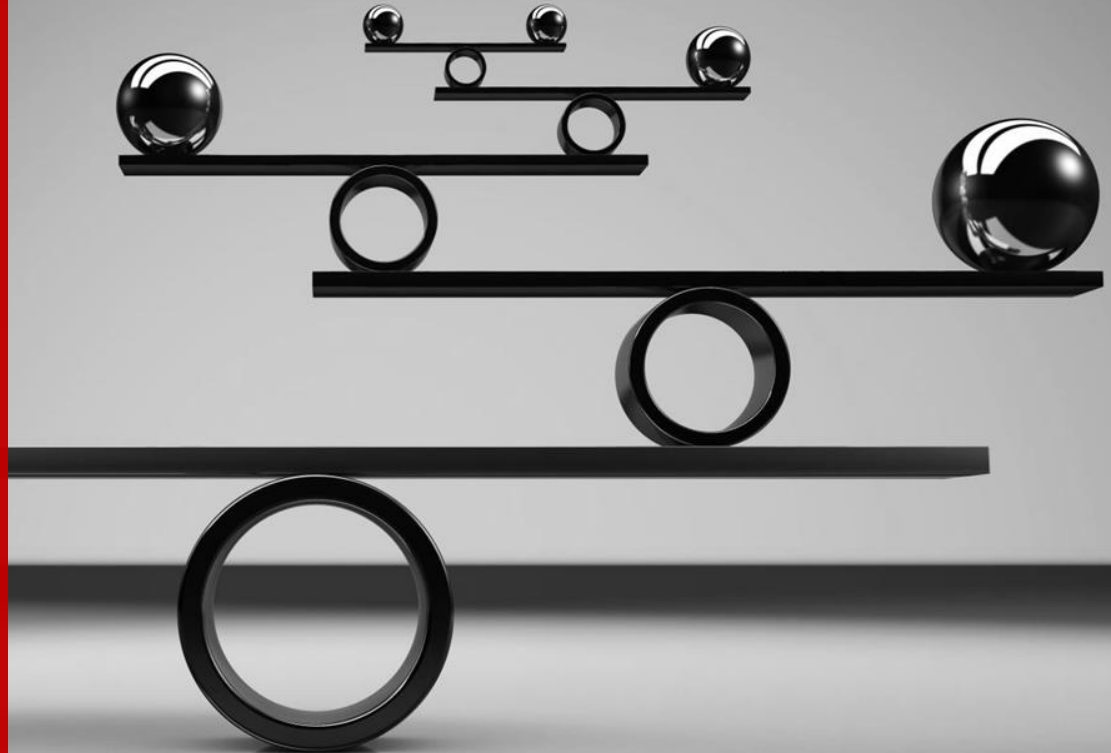
In the literature, several authors have proposed a method to perform an AI audit, but few have shown how to do it in practice.

Most papers focuses on the technical aspects, such as debiasing techniques and data set balancing, while the field of Value Sensitive Design is advocating for a more comprehensive approach.

Floridi et al. suggest a guide to implement the new AI Act requirements in its paper «CapAI».

The Grand Challenge in St. Gallen was a testbed to apply the AI Act.

## 2. BASIC PRINCIPLES OF AI ETHICS



# What is AI Ethics?

AI Ethics is a branch of Philosophy; it studies the ethical principles applicable to AI systems and addresses the implications, risks and problems derived from:

- 1) humans building and employing those systems, and
- 1) the behavior of autonomous systems

## Most famous ethics guidelines



Universal Guidelines for AI (2018)



OECD AI Principles (2019)



G20 AI Guidelines (2019)



UNESCO Recommendation on AI Ethics (2021)

Table 2 – Ethical principles identified in existing AI guidelines

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom & autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

Jobin, Anna, Marcello Lenca, and Effy Vayena. "The global landscape of AI ethics guidelines." *Nature Machine Intelligence* 1.9 (2019): 389-399.

# Trustworthy AI

The AI HLEG recognizes 3 keys elements of Trustworthy AI:



1. it should be **lawful**, complying with all applicable laws and regulations;



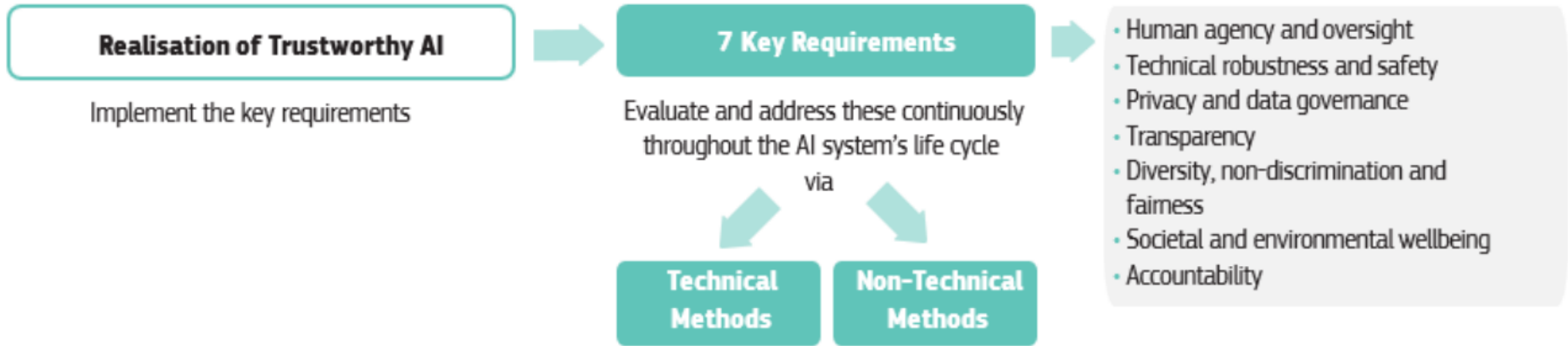
2. it should be **ethical**, ensuring adherence to ethical principles and values;  
and



3. it should be **robust**, both from a technical and social perspective, since, even with good intentions, AI systems can cause unintentional harm.

# The ALTAI List

AI HLEG has drawn a list of 7 elements:



Designed For		Date		Version	
<b>Scope</b> <ul style="list-style-type: none"> <li>What is this product designed for?</li> <li>In which context it operates?</li> </ul>	<b>Training Data</b> <ul style="list-style-type: none"> <li>How was the training data collected?</li> <li>How do you ensure its representativeness?</li> <li>Does your training dataset contain personal data?</li> <li>Who annotates the data and how quality is controlled?</li> <li>What is the data labeling process that you employ?</li> </ul>	<b>Algorithms &amp; Source Code</b> <ul style="list-style-type: none"> <li>Do you use open or proprietary sources? Why? Which?</li> <li>Who in the team is setting the heuristics (rules) which influence the output?</li> <li>How do you ensure the quality of used third-party codebases?</li> <li>What is your process of making the key architectural choices?</li> </ul>	<b>Decision Space</b> <ul style="list-style-type: none"> <li>What exactly does the product do?</li> <li>Can you provide the list of all possible outputs?</li> <li>How incorrectly supplied inputs are spotted?</li> <li>Is there anomaly detection in place?</li> </ul>		
<b>Users</b> <ul style="list-style-type: none"> <li>What type of users does this product have? (customers/admins/ etc)</li> <li>What are their roles?</li> </ul>	<b>Personal Data Processing</b> <ul style="list-style-type: none"> <li>Which personal data is collected by the product?</li> <li>What is the purpose of collecting personal data?</li> <li>How is this data processed? Used? Stored? Deleted?</li> </ul>	<b>Components &amp; Subprocessing</b> <ul style="list-style-type: none"> <li>Which third parties are engaged by the product?</li> <li>How do you evaluate the potential impacts of API on the quality of your product's output?</li> <li>How do you check the reliability of your data processing contractors?</li> </ul>	<b>Failure Modes</b> <ul style="list-style-type: none"> <li>How failures are detected and monitored?</li> <li>What are the possible failures of a product?</li> <li>What actions are performed if a product fails?</li> </ul>		
<b>Key Stakeholders</b> <ul style="list-style-type: none"> <li>Who are the key stakeholders?</li> <li>What influence do they have over the product?</li> <li>How do stakeholders interact with each other?</li> <li>How is the power distributed?</li> </ul>			<b>Values &amp; Interests</b> <ul style="list-style-type: none"> <li>What values do stakeholders/users have?</li> <li>Where these values can clash or create tensions?</li> <li>What is known at the moment and how assumptions are tested?</li> <li>How can you align your technology to the values you want to support/people desire?</li> </ul>	<b>Human in the Loop (HITL)</b> <ul style="list-style-type: none"> <li>What is the role of a human agent in the validation/verification of the outputs?</li> <li>What is the role of a human agent in refining the model performance?</li> <li>What is the decision-making power assigned to human agents responsible for the quality of output?</li> </ul>	<b>Model Performance Metrics</b> <ul style="list-style-type: none"> <li>Which metrics are used to evaluate the product performance?</li> <li>Which measures are used to re-evaluate Accuracy, Recall, Precision, and F1-Score?</li> </ul>
<b>Impact Assessment</b> <ul style="list-style-type: none"> <li>What potential harms can your product cause? (loss of opportunity, discrimination, economic loss, social stigma, detriment, emotional distress, etc)?</li> <li>What are the risks of the product's failure?</li> <li>What impact product can cause if deployed at scale?</li> <li>How is the product influencing the existing markets?</li> </ul>	<b>Explainability</b> <ul style="list-style-type: none"> <li>How is interpretability defined for the system?</li> <li>What interpretability methods are used?</li> <li>What metrics are used in result interpretation?</li> <li>How interpretations of the output are communicated?</li> </ul>	<b>Regulatory Landscape</b> <ul style="list-style-type: none"> <li>What is the regulatory context in which the product operates?</li> <li>Is the model portable to other market verticals?</li> <li>What are the involved regulatory risks?</li> </ul>	<b>Decision Feedback &amp; Objection</b> <ul style="list-style-type: none"> <li>How does the product allow for structured feedback?</li> <li>How can the user challenge the application output?</li> <li>Which are the third parties involved in claims/objection resolution?</li> </ul>		
	<b>Changes in Behavior</b> <ul style="list-style-type: none"> <li>Do the automated decisions have significant legal or similar effects on the users/stakeholders?</li> <li>How the users may change their behavior after use?</li> <li>What are the potentials for power imbalance?</li> </ul>		<b>Group Interactions</b> <ul style="list-style-type: none"> <li>What are potential changes in group behavior?</li> <li>How is the product addressing group interests?</li> <li>What new groups could be born due to the product deployment at scale?</li> </ul>	<b>Mitigation</b> <ul style="list-style-type: none"> <li>How do you test for bias and fairness? What fairness definitions do you employ and why?</li> <li>Does your team reflect a diversity of opinions, backgrounds, and thoughts?</li> <li>Do you have a process for redress if people are harmed by the outputs?</li> <li>How fast can you shut down your product in production if it behaves badly?</li> <li>Who and how should be informed?</li> </ul>	<b>Comments</b>

# Open Ethics Canvas

This instrument is used by several companies in order to perform an ethical assessment of their AI systems

**Green:** non-technical ethics

**Grey:** technical ethical assessment

**Yellow:** accountability and human oversight

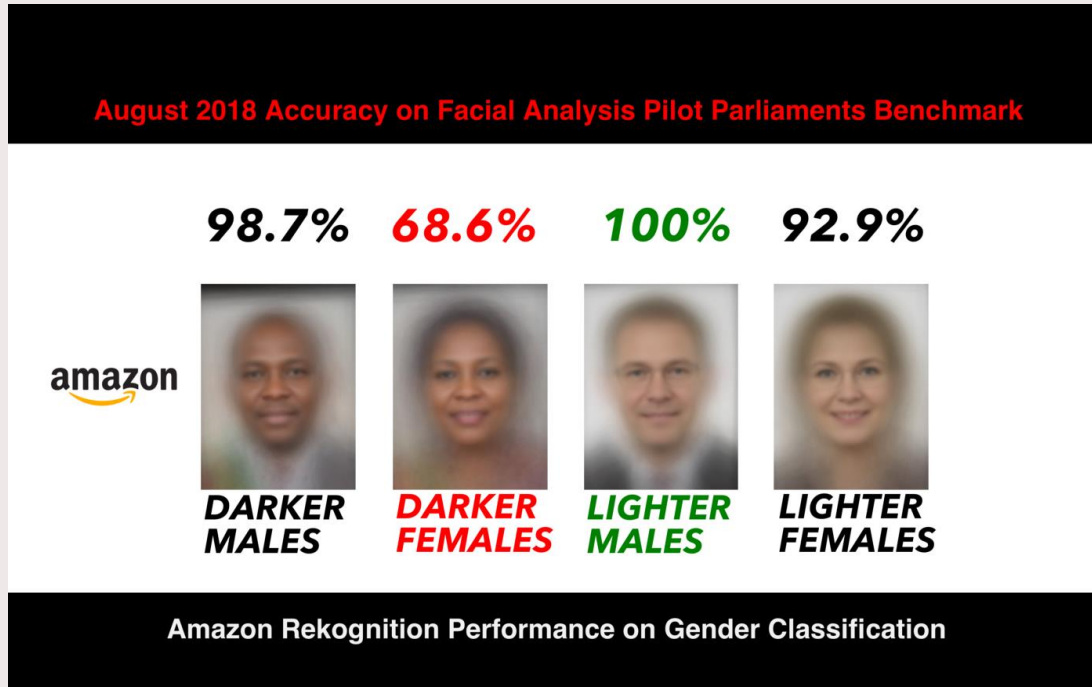
**Red:** legal assessment and societal liability

# What is bias?

The term **bias** is used by Friedman and Nissenbaum, who say it “refers to computer systems that **systematically and unfairly discriminates against certain individuals or groups of individuals in favor of others**”. By extension pre-existing **societal bias** is “**bias that originates from society at large, such as from organizations (e.g., industry), institutions (e.g., legal systems), or culture at large.** (E.g., gender biases present in the larger society that lead to the development of educational software that overall appeals more to boys than girls)”. Pre-existing bias can be implicit and subconscious, rather than malicious in intent.

Source: “Does Technology Have Race?”, Hankerson et al, 2016

# Examples

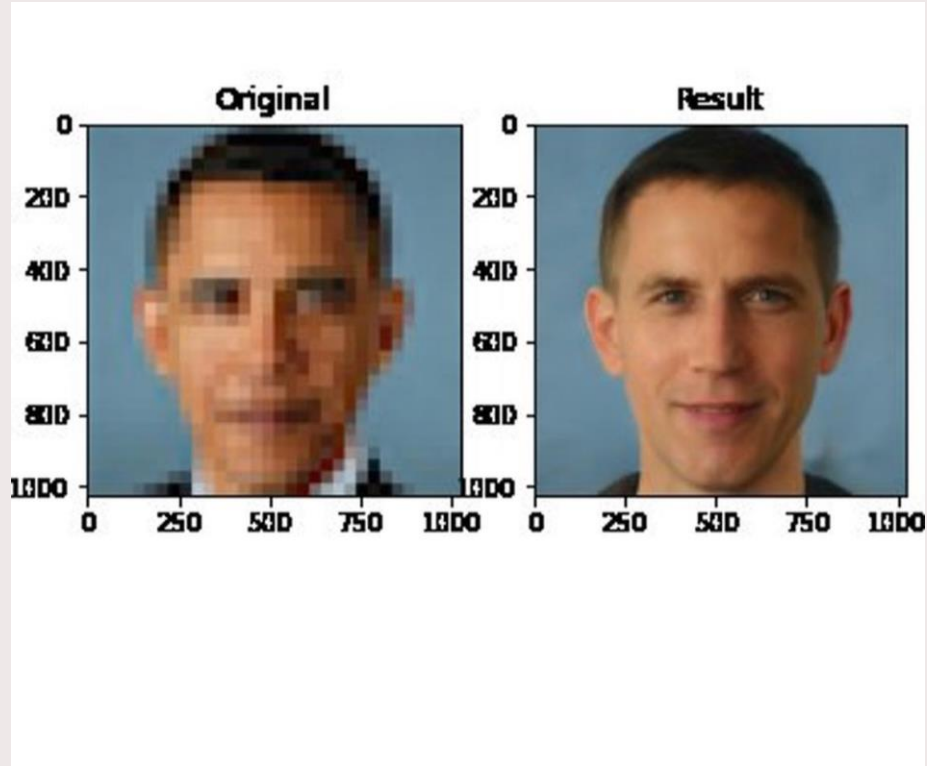


# Examples

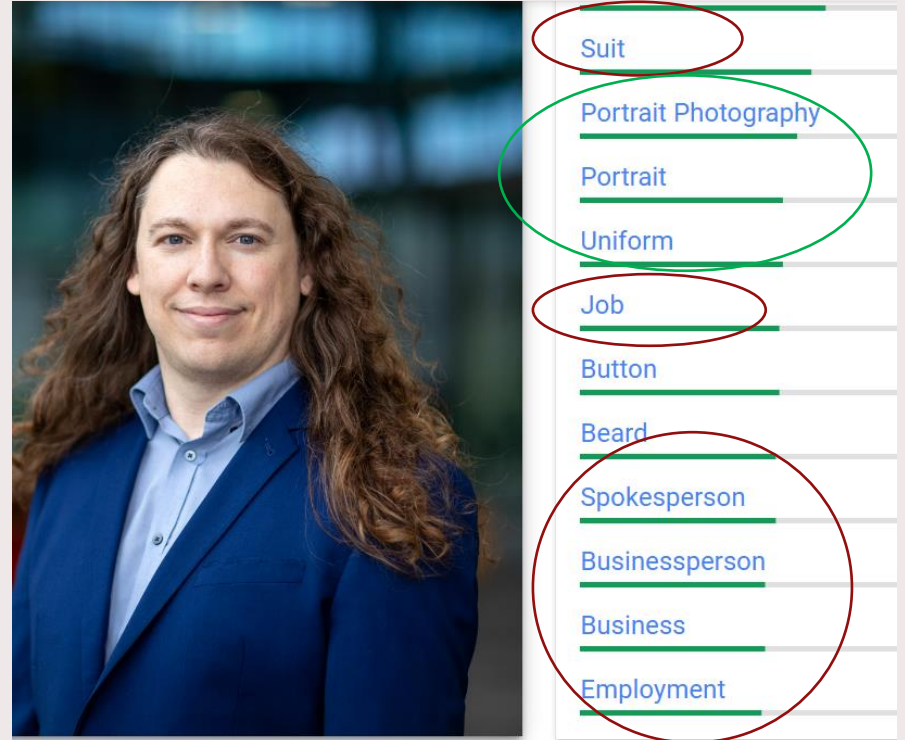
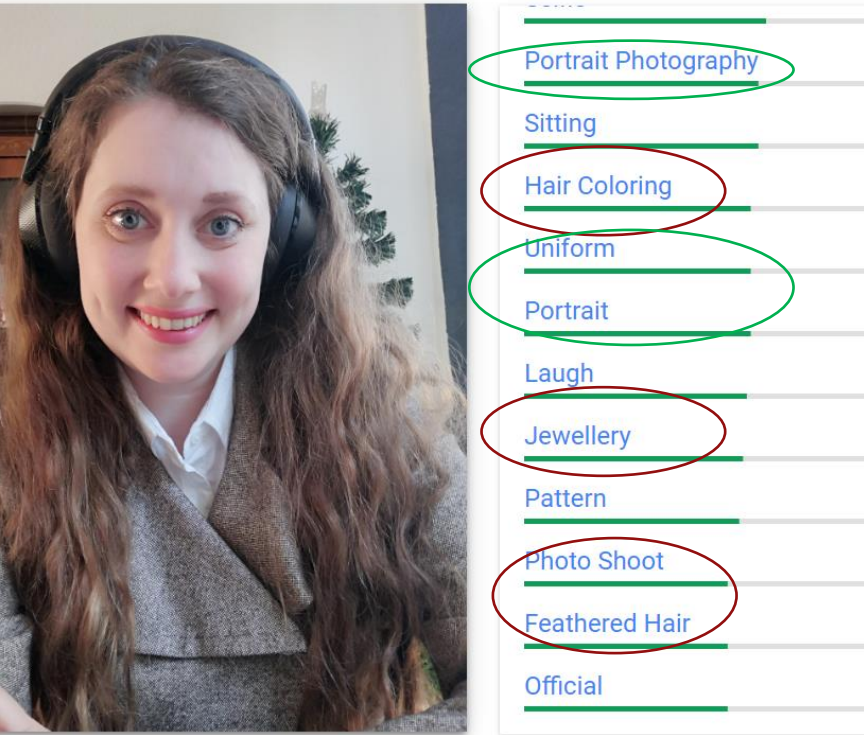


Google and its gorillas

# Examples



# Examples



### **3. THE NEW AI ACT PROPOSAL**



## What is considered AI in the new regulation?



- a) **Machine learning approaches**, including supervised, unsupervised and reinforcement learning, using a wide variety of methods including deep learning
- b) **Logic- and knowledge-based approaches**, including knowledge representation, inductive (logic) programming, knowledge bases, inference and deductive engines, (symbolic) reasoning and expert systems
- c) **Statistical approaches**, Bayesian estimation, search and optimization methods

# The protection of citizens' rights

The most important innovation of the proposal is the establishment of **four risks categories for AI systems**, in order to protect citizens' fundamental rights. The explanatory memorandum attached to the proposal, in fact, notes that the use of AI with its specific characteristics (e.g. opacity, complexity, dependency on data, autonomous behaviour) can adversely affect a number of fundamental rights enshrined in the EU Charter of Fundamental Rights. The proposal seeks to ensure a high level of protection for those fundamental rights and aims to address various sources of risks through a clearly defined **risk-based approach**.

# Classification of AI systems

The risk categories are related to the degree (intensity and scope) of risk for the safety or fundamental rights of citizens and are classified in **four** different groups:

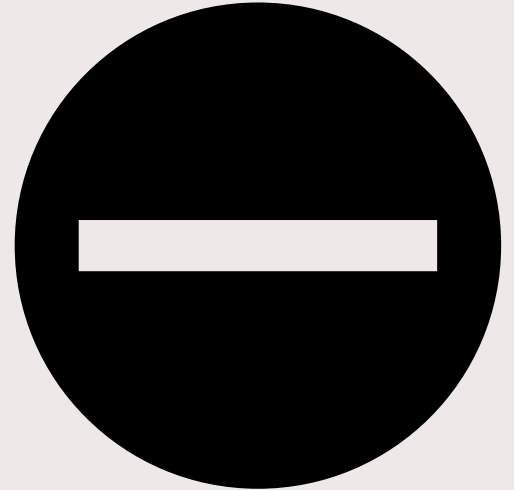
- (i) unacceptable risk**
- (ii) high risk**
- (iii) low risk**
- (iv) minimal risk**

Taking inspiration from the product safety legislation, the classification of risks is based on the intended purpose and modalities for which the AI system is used, not only on their specific function.

## Banned technology

Some AI systems are no longer allowed in the EU, with few exceptions, such as those that:

- deploy **subliminal techniques** beyond a person's consciousness in order to materially **distort** a person's **behaviour**
- **exploit** any of the **vulnerabilities** of a specific group of persons due to their age, physical or mental disability
- Are used by the P.A. for the **evaluation** or classification of the **trustworthiness** of natural persons based on their social behaviour or their characteristics
- use **'real-time' remote biometric identification** systems in **publicly accessible spaces** for the purpose of **law enforcement**



## The concept of 'High Risk'

Some AI system are considered **high risk** and have to meet certain requirements before being released to the market or to the public. Some examples includes systems used in:

- Machinery (including robots)
- Toys
- Medical devices
- In Vitro Diagnostic
- Vehicles
- Aircrafts
- Rails
- Health or life insurance



AI Regulation must be coordinated with MDR, GDPR, and other laws (e.g., civil liability rules)

## Other High Risk Systems

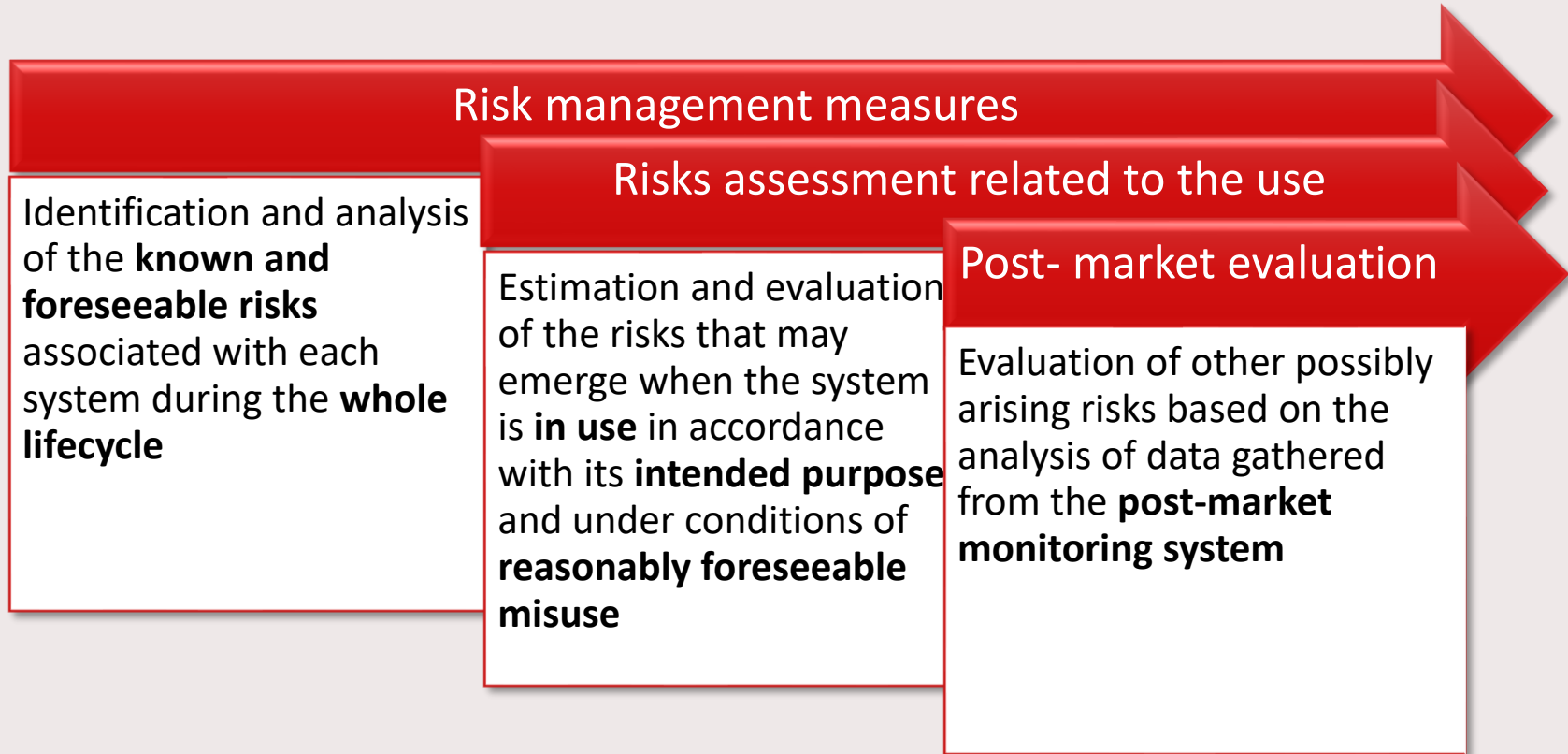
- Biometric identification and categorisation of natural persons
- Management and operation of critical infrastructure
- Education and vocational training
- Employment, workers management and access to self-employment
- Access to and enjoyment of essential private services and public services and benefits
- Law enforcement
- Migration, asylum and border control management
- Administration of Justice

# Mandatory requirements for High Risk Systems

## 1. Risk management system:

- a) identification and analysis of the **known and foreseeable risks** associated with each system
- b) estimation and evaluation of the risks that may emerge when the system is used in accordance with its **intended purpose** and under conditions of **reasonably foreseeable misuse**
- c) evaluation of other possibly arising risks based on the analysis of data gathered from the **post-market monitoring system**
- d) adoption of **suitable risk management measures**

# 1. Risk Management System

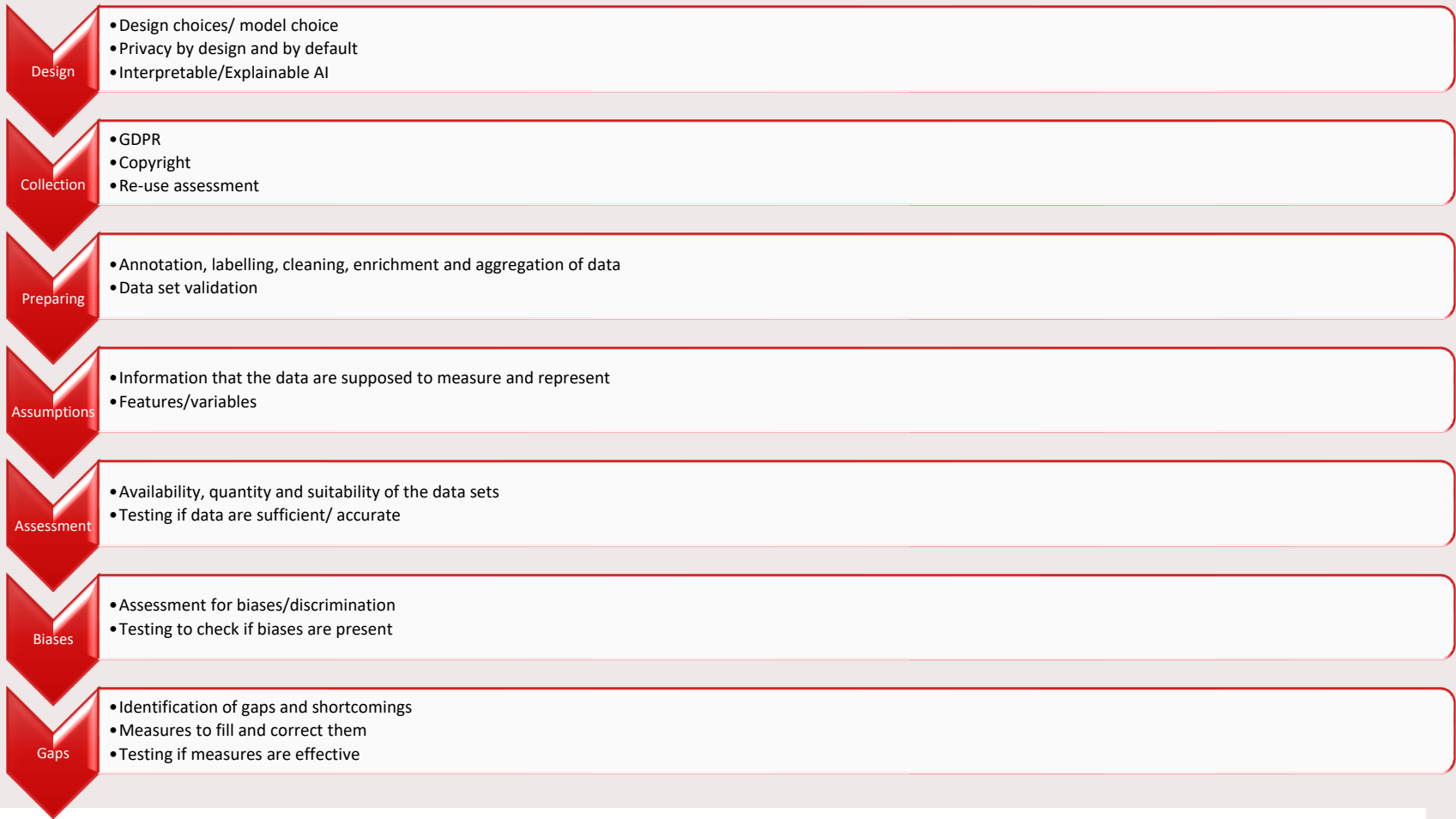


# Mandatory requirements for High Risk Systems

## 2. Data and data governance:

### Training, validation and testing data sets concerning:

- (a) the relevant **design choices**
- (b) **data collection**
- (c) relevant **data preparation processing operations**, such as annotation, labelling, cleaning, enrichment and aggregation
- (d) the formulation of **relevant assumptions**, notably with respect to the information that the data are supposed to measure and represent
- (e) a prior assessment of the **availability, quantity and suitability of the data sets** that are needed
- (f) examination in view of possible **biases**
- (g) the identification of any possible **data gaps or shortcomings**, and how those gaps and shortcomings can be addressed



# Mandatory requirements for High Risk Systems

## 3. **Technical documentation:**

Drawn **before releasing** the system

**Demonstrate compliance** with the requirements set out in the Regulation and provide national competent authorities and notified bodies with all the necessary information to assess the compliance with those requirements

Enabling the **automatic recording of events** ('logs') while the systems is operating

# Mandatory requirements for High Risk Systems

## 5. **Transparency** and provision of **information** to users:

- (a) the **identity and the contact** details of the provider and, where applicable, of its authorised representative;
- (b) the **characteristics, capabilities and limitations** of performance of the system
- (c) the **changes** to the system and its performance which have been pre-determined by the provider at the moment of the initial conformity assessment, if any;
- (d) the **human oversight measures**, including the technical measures put in place to facilitate the interpretation of the outputs of systems by the users;
- (e) the **expected lifetime** of the system and any **necessary maintenance** and care measures to ensure the proper functioning of that AI system, including as regards **software updates**

# Mandatory requirements for High Risk Systems

## 6. Human oversight

Through measures:

- (a) identified and **built into the system**, when technically feasible, by the provider **before** it is placed on the market or put into service, and/or
- (b) identified by the provider **before** placing the system on the market or putting it into service and that are appropriate **to be implemented by the user**

# Mandatory requirements for High Risk Systems

## 7. Accuracy, robustness and cybersecurity:

Systems shall be **resilient as regards errors, faults or inconsistencies** that may occur within the system or the environment in which the system operates

The robustness of systems may be achieved through **technical redundancy solutions**, which may include backup or fail-safe plans

Systems shall be **resilient as regards attempts by unauthorised third parties** to alter their use or performance by exploiting the system vulnerabilities

The technical solutions to address AI specific vulnerabilities shall include, where appropriate, measures to **prevent and control for attacks** trying to manipulate the training dataset ('data poisoning'), inputs designed to cause the model to make a mistake ('adversarial examples'), or model flaws

# Mandatory requirements for High Risk Systems

## 8. Quality management system/1:

- a) a strategy for **regulatory compliance**
- b) techniques, procedures and systematic actions for the **design, design control and design verification of the system**
- c) techniques, procedures and systematic actions for the **development, quality control and quality assurance** of the system
- d) **examination, test and validation procedures**
- e) **technical specifications**, including standards, to be applied
- f) systems and procedures for **data management**, including data collection, data analysis, data labelling, data storage, data filtration, data mining, data aggregation, data retention and any other operation regarding the data that is performed **before** - and for the purposes of - the placing on the market or putting into service of the systems (see n. 2)



# Mandatory requirements for High Risk Systems

## 8. Quality management system/2:

- g) the setting-up, implementation and maintenance of a **post-market monitoring system**
- h) procedures related to the **reporting of serious incidents and malfunctioning**
- i) the handling of **communication with competent authorities**, including sectoral ones, **providing or supporting the access to data**, notified bodies, other operators, customers or other interested parties
- j) systems and procedures for **record keeping** of all relevant documentation and information
- k) **resource management**, including security of supply related measures
- l) an **accountability framework** setting out the responsibilities of the management and other staff with regard to all aspects listed in this paragraph

## 9. Providers shall draw up technical documentation required by the Regulation

## Transparency obligations for certain AI systems



Providers shall ensure that AI systems **intended to interact with natural persons** are designed and developed in such a way that natural persons are **informed that they are interacting with an AI system**, unless this is obvious from the circumstances and the context of the use

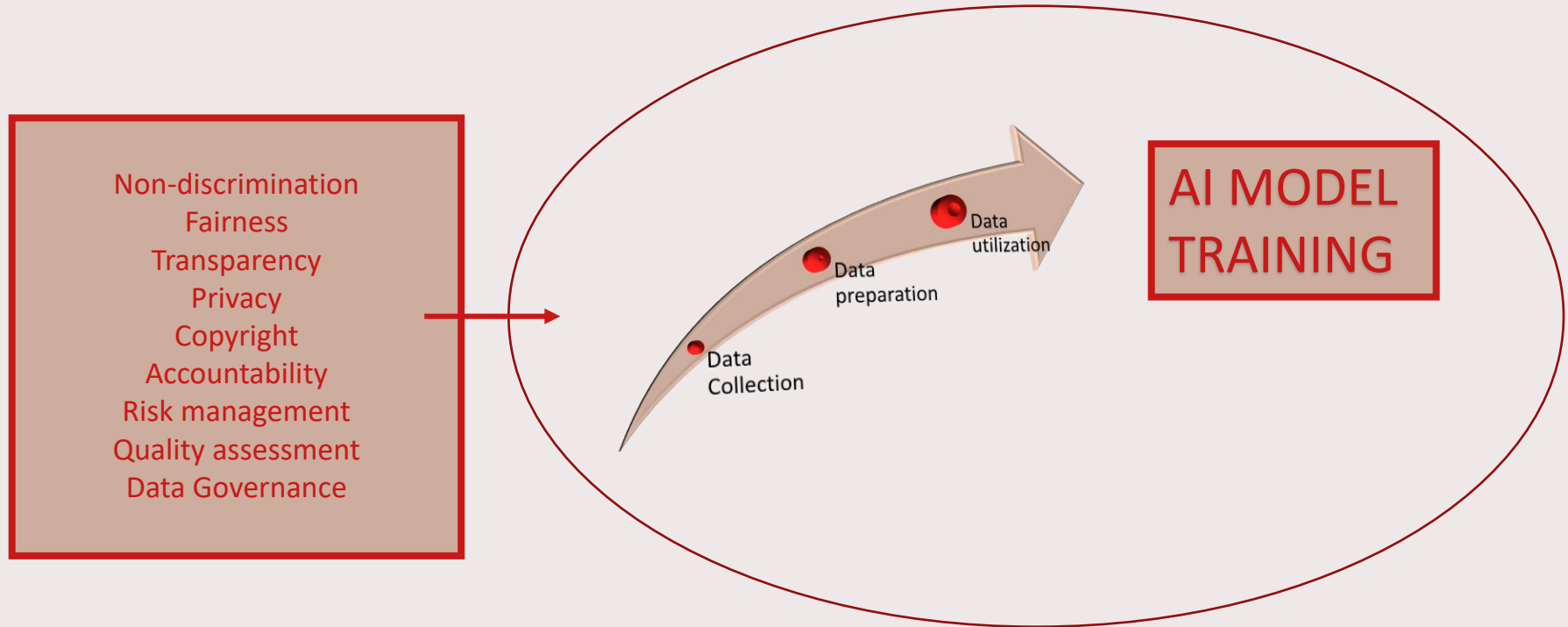


Users of an AI system that **generates or manipulates image, audio or video content** that appreciably **resembles existing persons, objects, places or other entities or events** and would falsely appear to a person to be authentic or truthful ('**deep fake**'), shall **disclose** that the content has been artificially generated or manipulated

**4. APPLYING  
ETHICAL  
PRINCIPLES TO  
DATA SETS:  
PRACTICAL  
EXAMPLES**



# Applying ethical and legal principles in practice



# DATA SET FAIRNESS 1

## What is data set fairness?

Research has shown how **data sets convey social bias** in AI systems, especially those based on machine learning. A biased data set is **not representative of reality** and might contribute to perpetuate societal biases within the model. To tackle this problem, it is important to understand how to **avoid biases, errors, and unethical practices** while **creating the data sets**.

# DATA SET FAIRNESS 2

## What is data set fairness?

Mainstream works in computer science addresses the problem of fairness in a context of prediction-based decisions. Usually, mathematical definitions requires **achieving parities across distributions of outcomes or scores** (common examples include statistical parity and calibration). Definitions can refer to **individuals** (e.g., requiring that similar individuals are treated similarly) or **groups** (e.g., requiring that any error measure is equal across groups, defined by so-called protected characteristics like gender or race).

Mathematical definitions can reflect philosophical conceptualisations, such as the **equality of outcomes and the equality of opportunities**. In general, computer science literature expresses in empirical and formal terms philosophical problems of **distributive justice** in which the goal of fairness is to achieve an **equal allocation of goods**.

# DATA SET FAIRNESS 3

## What is data set fairness?

The concept of fairness is frequently found also in **text of law** and it has been extensively analysed by the legal scholarship, having its roots in the "**bona fide**" **concept in Roman law**. The fairness principle is an **overarching obligation** in the European Digital Strategy, for example in Article 5, par. 1, of GDPR, and in the Data Act, but it is also found in the EU and in other countries in many pieces of law, such as Article 8, par. 2, of the Charter of Fundamental Rights of the European Union, contract law, and others.

The data protection law scholarship divide the fairness principle into two elements, **procedural fairness** and **fair balancing**, noting that the concept of fairness is related to the idea of **vulnerability**. The GDPR aims at protecting data subjects, who are in a vulnerable position as opposed to data controllers, therefore it tries to make a balance between the fundamental rights and freedoms of the former and the interests of the latter.

# FAIRNESS ASSESSMENT OF DATA SETS

So far, the assessment of algorithmic fairness in the context of data has followed **two main directions**. A research track regards the analysis of **data curation processes** or the development of new **documentation frameworks**. Recently, Fabris et. Al annotated hundreds of data sets employed in algorithmic research efforts specifying key fairness-related characteristics (e.g. the domain of application, the target task and the sensitive attributes encoded in the data set) in order to address the so-called “documentation debt”.

Another direction focuses on the **empirical analysis of fairness-related data dimensions**. These efforts include **pre-processing techniques** such as sanitization and instance re-weighting and procedures evaluating distinct sources of discrimination in data collection or studying the effect of data-related factors on the fairness of algorithms.

# OUR WORK IN TRIESTE

## Investigating Semi-Automatic Assessment of Data-set Fairness by Means of Fuzzy Logic

Chiara Gallese

*Department of Mathematics and Geosciences  
University of Trieste  
Trieste, Italy  
email address or ORCID*

Teresa Scantanburlo

*Department of Environmental Sciences, Informatics and Statistics  
Ca' Foscari University of Venice  
University of Trieste  
Trieste, Italy*

Luca Manzoni

*Department of Mathematics and Geosciences  
University of Trieste  
Trieste, Italy  
email address or ORCID*

Marco S. Nobile

*Department of Environmental Sciences, Informatics and Statistics  
Ca' Foscari University of Venice  
Venice, Italy  
Department of Industrial Engineering & Innovation Sciences  
Eindhoven University of Technology  
Eindhoven, The Netherlands  
Bicocca Bioinformatics, Biostatistics and  
Bioimaging research center (B4)  
Monza, Italy  
marco.nobile@unive.it*

# FANFAIR AIMS

Our work moves along the line of empirical research with some important distinctions. The fairness-related features we explore are **agnostic about the sensitive attribute** in that they capture essential vulnerabilities that may contribute to model (un)fairness.

Another distinct aspect of our approach is to integrate **statistical information** about the data sets (**balance, numerosity, unevenness, incompleteness**) with **qualitative considerations (compliance, quality)** which take into account legal and ethical requirements of data-related practices (e.g. GDPR, principles of research integrity). This integration is operated through the application of **fuzzy rule-base logic** and reflected in a score that can be indicative of the fairness potentially introduced by a given data set.

## FANFAIR DEFINITION OF FAIRNESS

Although in our paper we take into consideration GDPR, we do not limit the concept of (legal) fairness to data protection law, but we expand the notion to other areas of law, including the **non-discrimination principle**, that is found, inter alia, in the European Convention of Human Rights, the **equity principle**, and the **bona fide principle** in contractual agreements.

In our opinion, a data set which data collection has been flawed by discrimination or another breach of the law (e.g., copyright law, criminal law, medical law) cannot be considered fair.

The **lawfulness criteria** is also in line with the High-Level Expert Group on AI (HLEG AI)'s guidelines on Trustworthy AI.

# FANFAIR FUZZY LOGIC ASSESSMENT

In order to assess the potential fairness of a data set, we identified six different elements that should be addressed before employing the data set. To identify each criteria, we were inspired by general principle of civil law and EU law, and established ethics guidelines.

TABLE I  
DESCRIPTION OF THE DATA SET FEATURES / LINGUISTIC VARIABLES  
CONSIDERED IN THIS WORK.

Feature	Meaning
balance	How balanced is the data set, with respect to the output labels.
numerosity	The numerosity of instances with respect to the number of features.
evenness	A metric assessing the amount of outliers in the data set.
compliance	How much the data set is compliant with respect to governmental rules.
quality	A user-defined evaluation of the overall quality of the data set.
incompleteness	This metric measures the amount of missing data in the data set.

# RESULTS OF FANFAIR

These variables represent important characteristics that are automatically extracted by the data set, or (in the case of compliance and quality) are entered by the user. FANFAIR returns a summary of the contribution of all variables to the final fairness value. The contributions are expressed both in fuzzy terms (e.g., “compliance is fairly high”) and by using a visual plot. The final fairness value is calculated as a value between 0 (totally unfair) and 1 (maximum theoretical fairness).

TABLE II  
RULE-BASE USED BY FANFAIR FOR THE ASSESSMENT OF DATA SET FAIRNESS.

Rule 1:	IF balance IS high THEN $\varphi$ IS high_fairness
Rule 2:	IF balance IS low THEN $\varphi$ IS low_fairness
Rule 3:	IF numerosity IS high THEN $\varphi$ IS high_fairness
Rule 4:	IF numerosity IS low THEN $\varphi$ IS low_fairness
Rule 5:	IF unevenness IS high THEN $\varphi$ IS low_fairness
Rule 6:	IF unevenness IS low THEN $\varphi$ IS high_fairness
Rule 7:	IF compliance IS high THEN $\varphi$ IS high_fairness
Rule 8:	IF compliance IS low THEN $\varphi$ IS low_fairness
Rule 9:	IF quality IS high THEN $\varphi$ IS high_fairness
Rule 10:	IF quality IS low THEN $\varphi$ IS low_fairness
Rule 11:	IF incompleteness IS high THEN $\varphi$ IS low_fairness
Rule 12:	IF incompleteness IS low THEN $\varphi$ IS high_fairness

# Our pilot project in Eindhoven

RAISE project – a collaboration with NXP semiconductors



## Developing AI models for hearing aids

We are working on a project to develop an AI model to help people with **hearing impairments**

The goal is to **help them hear better and understand what it is said in a conversation**, so the model needs to be able to recognize all kind of voices in all settings

People with hearing impairments have **the right not to be excluded** in a conversation

People with speech impairments, with a different accent or pronunciation, with a low/high tone of voice, or that are slow in talking, have **the right to have their voice recognized**

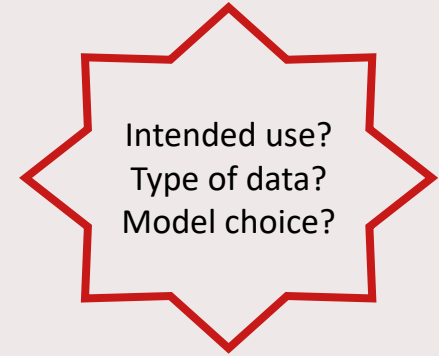
# What could go wrong?

If the data set is **biased** and different categories of people are not represented, this could lead to **discrimination** towards both people with hearing impairments and **underrepresented groups**, as the model will **not be able to recognize all voices**



# Mozilla Common Voice Data Set

SIZE	2 GB
VERSION	nI_105h_2022-01-19
VALIDATED HR. TOTAL	98
OVERALL HR. TOTAL	105
LICENSE	CC-0
NUMBER OF VOICES	1,462
AUDIO FORMAT	MP3
SPLITS	<small>Age</small> 22% 19 - 29, 15% 40 - 49, 10% 30 - 39, 8% 50 - 59, 2% < 19, 1% 60 - 69 47% Male, 11% Female



Validation info



Type of license



Number of examples



Age and gender composition



# Terms & Conditions

\*Email  
c.g.gallese.nobile@tue.nl

You are prepared to initiate a download of **2 GB**

**You agree** to not attempt to determine the identity of speakers in the Common Voice dataset

[Download Dutch](#) 

sha256 checksum: a173515c0f7c99559ef5a446c26d0c16e91d72201ffd64bd68703b7776fee2c1



Privacy

# Looking into the data set metadata

	A	B	C	D	E	F	G	H	I	J
1	client_id	path	sentence	up_votes	down_votes	age	gender	accents	locale	segment
2	0d709133bf209da7f0164653b6e5f9e	common_voice_nl_17699535.mp3	De Aboriginals zijn de oorspronkelijke bewoners van AustraliA	2	0	fifties	male	Nederland	nl	
3	0fca93407be6d482019f2463e60fbaf	common_voice_nl_17694848.mp3	Mijn toetsenbord zit vol stof.	2	1				nl	
4	175d4117110538cc68a8a0157a7f0a1	common_voice_nl_18441136.mp3	Ze had de bank beschadigd met haar skateboard.	2	0				nl	
5	29ecd9ca5970fd1bc1297bb4a9a5f3e	common_voice_nl_30205425.mp3	Deze top heeft niet in alle opzichten aan de verwachtingen vo	2	0				nl	
6	3c448c76f72ee15009da6c9e474dc9f	common_voice_nl_19966492.mp3	Waar laat jij je onderhoud doen?	2	0	thirties	male		nl	
7	452cfa9d7e49e72eac2bb7ff3d5f79c	common_voice_nl_17700547.mp3	Na het lezen van vele beoordelingen had ze eindelijk haar oog	2	1	fifties	male	Nederland	nl	
8	4ac8a9187cfd11da401d6829199c18	common_voice_nl_21571581.mp3	De tampons zijn op.	2	0	twenties	male	Belgisch N	nl	
9	4d6898ec762a9590a973dc27579b93	common_voice_nl_25310936.mp3	Proberen wij kooldioxide-emissies te beperken of proberen wi	2	0				nl	
10	81251914ca5d3e3eb5de368076aecc	common_voice_nl_21431804.mp3	Marijke kent Olivier nu al meer dan twee jaar.	2	0	twenties	female	Nederland	nl	
11	5f1e2fabc5a5c63d880ce655697814e	common_voice_nl_28458010.mp3	Mijn partij vindt dat onverkwikelijk.	2	0				nl	
12	723a65c399699beeb3c730e141775f	common_voice_nl_19031796.mp3	Het voeren van brood aan eenden is eigenlijk ongezond voor c	2	0				nl	
13	77fc542b71acf6510c80e9a681c128f	common_voice_nl_30125083.mp3	Juni dit jaar is veel te laat.	2	0	seventies	male		nl	
14	79d570bd61fb0fc4d363d0a1e044cf	common_voice_nl_17695742.mp3	Parket moet je stofzuigen, tegels moet je dweilen.	2	1				nl	
15	7aa3af7141d503c7c449404c3f79393	common_voice_nl_19691757.mp3	In onze buurt kent iedereen elkaar.	2	1	teens	male	Nederland	nl	
16	8d9fece2b05d9fe37e9c66e45767cae	common_voice_nl_22759350.mp3	Naar mijn oordeel valt er namelijk niet meer zoveel te verbete	2	0				nl	
17	9062e9bc9c6ee752e6f836d4296ee3	common_voice_nl_30057990.mp3	Daarom wordt voorgesteld dit soort pogingen te verbieden en	2	0				nl	
18	a01a8a30db997795dc07295a000bbc	common_voice_nl_17706390.mp3	Heb je de tafel gedekt?	2	1				nl	
19	a64ddf48b5b3b6700b146de59126a0	common_voice_nl_17696892.mp3	Ine lag in de strandstoel.	2	1				nl	
20	a9f1622fcaedf8cc1fba8d13c58564b1	common_voice_nl_19442483.mp3	Zeven plus vier is elf.	2	0				nl	
21	b51386637c01eace0c2291e27896d7	common_voice_nl_24722260.mp3	Het verslag staat echter bol van de neoliberale beleidsvoorste	2	0				nl	
22	bd3c4d0d793087205cd0adef921bd1	common_voice_nl_20177910.mp3	Ik denk dat ik gevolgd word.	2	1				nl	
23	bf115c7dcf62b2ac4dcc49baa4d035a	common_voice_nl_24733969.mp3	Wij verzetten ons tegen deze begrotingsposten.	2	0	fifties	female		nl	
24	bf9594e371b2f199760403701a1769	common_voice_nl_17703134.mp3	Ze hadden hun gevel bordeauxrood geschilderd.	2	1				nl	
25	c18bab9891bf6f1d34a29d6e0dbdbb	common_voice_nl_17700297.mp3	Hij ligt met zijn armen gekruist op de zetel.	2	0				nl	
26	c80d131260865d629ef044c0f4a25bc	common_voice_nl_29432749.mp3	Ook de afgelopen dagen verloren weer tientallen mensen het	2	0				nl	
27	d3441ac81d7899453a061a53222cf1	common_voice_nl_18294089.mp3	De bomen worden eerst gekapt, worden dan in de zagerij tot p	2	1				nl	
28	dcdb0f48072b0c136ddc02a09429e1	common_voice_nl_22499172.mp3	Het is dus een minimumconsensus.	2	0				nl	

# Looking into the data set metadata

	A	B	C	D	E	F	G	H	I	J
1	client_id	path	sentence	up_votes	down_votes	age	gender	accents	locale	segment
2	0d709133bf209da7f0164653b6e5f9e	common_voice_nl_17699535.mp3	De Aboriginals zijn de			0 fifties	male	Nederland	nl	
3	0fca93407be6d482019f2463e60fbaf	common_voice_nl_17694848.mp3	Mijn toetsenbord zit			1		nl		
4	175d4117110538cc68a8a0157a7f0a1	common_voice_nl_18441136.mp3	Ze had de bank besch			0		nl		
5	29ecd9ca5970fd1bc1297b74a9a5f3e	common_voice_nl_30205425.mp3	Deze top heeft niet i			0				
6	3c448c76f72ee15009da6c9e474dc9f	common_voice_nl_19966492.mp3	Waar laat jij je onder			0 thirties	male		nl	
7	452cfa9d7e49e72eac2bb7ff3d5f79c	common_voice_nl_17700547.mp3	Na het lezen van vele			1 fifties	male	Nederland	nl	
8	4ac8a9187cfd11da401d6829199c18	common_voice_nl_21571581.mp3	De tampons zijn op.			0 twenties	male	Belgisch N	nl	
9	4d6898ec762a9590a973dc27579b93	common_voice_nl_25310936.mp3	Proberen wij kooldio			0		nl		
10	581251914ca5d3eeb5de368076aecc	common_voice_nl_21431804.mp3	Marijke kent Olivier r			0 twenties	female	Nederland	nl	
11	5f1e2fab5a5c63d880ce655697814e	common_voice_nl_28458010.mp3	Mijn partij vindt dat e			0		nl		
12	723a65c399699beeb3c730e141775f	common_voice_nl_19031796.mp3	Het voeren van broo			0		nl		
13	77fc542b71ac6f510c80e9a681c128f	common_voice_nl_30125083.mp3	Juni dit jaar is veel te laat.	2		0 seventies	male		nl	
14	79d570bd61fb0fc4d363d0a1e044cf	common_voice_nl_17695742.mp3	Parket moet je stofzuigen, tegels moet je dweilen.	2		1			nl	
15	7aa3af7141d503c7c449404c3f79393	common_voice_nl_19691757.mp3	In onze buurt kent iedereen elkaar.	2		1 teens	male	Nederland	nl	
16	8d9fece2b05d9fe37e9c66e45767cae	common_voice_nl_22759350.mp3	Naar mijn oordeel valt er namelijk niet meer zoveel te verbete	2		0			nl	
17	9062e9bc9c6ee752e6f836d4296ee3	common_voice_nl_30057990.mp3	Daarom wordt voorgesteld dit soort pogingen te verbieden en	2		0			nl	
18	a01a8a30db997795dc07295a000bbc	common_voice_nl_17706390.mp3	Heb je de tafel gedekt?	2		1			nl	
19	a64ddf48b5b3b6700b146de59126a0	common_voice_nl_17696892.mp3	Ine lag in de strandstoel.	2		1			nl	
20	a9f1622fcaedf8cc1fba8d13c58564b1	common_voice_nl_19442483.mp3	Zeven plus vier is elf.	2		0			nl	
21	b51386637c01eace0c2291e27896d7	common_voice_nl_24722260.mp3	Het verslag staat echter bol van de neoliberale beleidsvoorste	2		0			nl	
22	bd3c4d0d793087205cd0adef921bd1	common_voice_nl_20177910.mp3	Ik denk dat ik gevolgd word.	2		1			nl	
23	bf115c7dcf62b2ac4dcc49baa4d035a	common_voice_nl_24733969.mp3	Wij verzetten ons tegen deze begrotingsposten.	2		0 fifties	female		nl	
24	bf9594e371b2f199760403701a1769	common_voice_nl_17703134.mp3	Ze hadden hun gevel bordeauxrood geschilderd.	2		1			nl	
25	c18bab9891bf6f1d34a29d6e0dbddb	common_voice_nl_17700297.mp3	Hij ligt met zijn armen gekruist op de zetel.	2		0			nl	
26	c80d131260865d629ef044c0f4a25bc	common_voice_nl_29432749.mp3	Ook de afgelopen dagen verloren weer tientallen mensen het	2		0			nl	
27	d3441ac81d7899453a061a53222cf1	common_voice_nl_18294089.mp3	De bomen worden eerst gekapt, worden dan in de zagerij tot p	2		1			nl	
28	dcdb0f48072b0c136ddc02a09429e1	common_voice_nl_22499172.mp3	Het is dus een minimumconsensus.	2		0			nl	

Are all relevant metadata present?  
How can you fill the missing data?

# Looking into the data set metadata

	A	B	C	D	E	F	G	H	I	J
1	client_id	path	sentence	up_votes	down_votes	age	gender	accents	locale	segment
2	0d709133bf209da7f0164653b6e5f9e	common_voice_nl_17699535.mp3	De Aboriginals zijn de oorspronkelijke bewoners van Australië.	2	0	fifties	male	Nederland	nl	
3	0fca93407be6d482019f2463e60fbaf	common_voice_nl_17694848.mp3	Mijn toetsenbord zit vol stof.	2	1				nl	
4	175d4117110538cc68a8a0157a7f0a1	common_voice_nl_18441136.mp3	Ze had de bank beschadigd met haar skateboard.	2	0				nl	
5	29ecd9ca5970fd1bc1297bb4a9a5f3e	common_voice_nl_30205425.mp3	Deze top heeft niet in alle opzichten aan de verwachtingen voldaan.	2	0				nl	
6	3c448c76f72ee15009da6c9e474dc9f	common_voice_nl_19966492.mp3	Waar laat jij je onderhoud doen?	2	0	thirties	male		nl	
7	452cfa9d7e49e72eac2bb7ff3d5f79c	common_voice_nl_17700547.mp3	Na het lezen van vele beoordelingen had ze eindelijk haar oog gevonden.	2	1	fifties	male	Nederland	nl	
8	4ac8a9187cfd11da401d6829199c18	common_voice_nl_21571581.mp3	De tampons zijn op.	2	0	twenties	male	Belgisch N	nl	
9	4d6898ec762a9590a973dc27579b93	common_voice_nl_25310936.mp3	Proberen wij kooldioxide-emissies te beperken of proberen wij ze te verminderen?	2	0				nl	
10	581251914ca5d3eeb5de368076aecc	common_voice_nl_21431804.mp3	Marijke kent Olivier nu al meer dan twee jaar.	2	0	twenties	female	Nederland	nl	
11	5f1e2fabcb5a5c63d880ce655697814e	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	0				nl	
12	723a65c399699beeb3c730e141775f	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	0				nl	
13	77fc542b71acf6510c80e9a681c128f	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	0	seventies	male		nl	
14	79d570bd61fb0fcf4d363d0a1e044cf	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	1				nl	
15	7aa3af7141d503c7c449404c3f79393	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	1	teens	male	Nederland	nl	
16	8d9fce2b05d9fe37e9c66e45767cae	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	0				nl	
17	9062e9bc9c6ee752e6f836d4296ee3	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	0				nl	
18	a01a8a30db997795dc07295a000bbc	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	1				nl	
19	a64ddf48b5b3b6700b146de59126a0	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	1				nl	
20	a9f1622fcaedf8cc1fba8d13c58564bl	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	0				nl	
21	b51386637c01eace0c2291e27896d7	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	0				nl	
22	bd3c4d0d793087205cd0adef921bd1	common_voice_nl_17700547.mp3	Ik heb een beetje moeite met lezen.	2	1				nl	
23	bf115c7dcf62b2ac4dcc49baa4d035a	common_voice_nl_24733969.mp3	Wij verzetten ons tegen deze begrotingsposten.	2	0	fifties	female		nl	
24	bf9594e371b2f199760403701a1769	common_voice_nl_17703134.mp3	Ze hadden hun gevel bordeauxrood geschilderd.	2	1				nl	
25	c18bab9891bf6f1d34a29d6e0dbdbb	common_voice_nl_17700297.mp3	Hij ligt met zijn armen gekruist op de zetel.	2	0				nl	
26	c80d131260865d629ef044c0f4a25bc	common_voice_nl_29432749.mp3	Ook de afgelopen dagen verloren weer tientallen mensen het leven.	2	0				nl	
27	d3441ac81d7899453a061a53222cf1	common_voice_nl_18294089.mp3	De bomen worden eerst gekapt, worden dan in de zagerij tot splinters vermalen.	2	1				nl	
28	dcdb0f48072b0c136ddc02a09429e1	common_voice_nl_22499172.mp3	Het is dus een minimumconsensus.	2	0				nl	

**Individual identifier**  
=  
**Not anonymous**  
**Beware of linking risk**

# Looking into the data set metadata

	A	B	C	D	E	F	G	H	I	J
1	client_id	path	sentence	up_votes	down_votes	age	gender	accents	locale	segment
2	0d709133bf209da7f0164653b6e5f9e	common_voice_nl_17699535.mp3	De Aboriginals zijn de oorspronkelijke bewoners van Australi	2	0	fifties	male	Nederland	nl	
3	0fca93407be6d482019f2463e60fbaf	common_voice_nl_17694848.mp3	Mijn toetsenbord zit vol stof.	2	1				nl	
4	175d4117110538cc68a8a0157a7f0a1	common_voice_nl_18441136.mp3	Ze had de bank beschadigd met haar skateboard.	2	0				nl	
5	29ecd9ca5970fd1bc1297bb4a9a5f3e	common_voice_nl_30205425.mp3	Deze top heeft niet in alle opzichten aan de verwachtingen vo	2	0				nl	
6	3c448c76f72ee15009da6c9e474dc9f	common_voice_nl_19966492.mp3	Waar laat jij je onderhoud doen?	2	0	thirties	male		nl	
7	452cfa9d7e49e72eac2bb7ff3d5f79c	common_voice_nl_17700547.mp3	Na het lezen van vele beoordelingen had ze eindelijk haar oog	2	1	fifties	male	Nederland	nl	
8	4ac8a9187cfd11da401d6829199c18	common_voice_nl_21571581.mp3	De tampons zijn op.	2	0	twenties	male	Belgisch N	nl	
9	4d6898ec762a9590a973dc27579b93	common_voice_nl_25310936.mp3	Proberen wij kooldioxide-emissies te beperken of proberen wi	2	0				nl	
10	581251914ca5d3e3eb5de368076aecc	common_voice_nl_21431804.mp3	Marijke kent Olivier nu al meer dan twee jaar.	2	0	twenties	female	Nederland	nl	
11	5f1e2fabc5a5c63d880ce655697814e	common_voice_nl_28458010.mp3	Mijn partij vindt dat onoverkwankelijk.	2	0				nl	
12	723a65c399699b3c730e141775f	common_voice_nl_1	zond voor c	2	0				nl	
13	77fc542b71acf6510c80e9a681c128f	common_voice_nl_3		2	0	seventies	male		nl	
14	79d570bd61fb0fc4d363d0a1e044cf	common_voice_nl_1		2	1				nl	
15	7aa3af7141d503c7c449404c3f79393	common_voice_nl_1		2	1	teens	male	Nederland	nl	
16	8d9fce2b05d9fe37e9c66e45767cae	common_voice_nl_2	l te ve bete	2	0				nl	
17	9062e9bc9c6ee752e6f836d4296ee3	common_voice_nl_3	erbieden en	2	0				nl	
18	a01a8a30db997795dc07295a000bbc	common_voice_nl_1		2	1				nl	
19	a64ddf48b5b3b6700b146de59126a0	common_voice_nl_1		2	1				nl	
20	a9f1622fcaedf8cc1fba8d13c58564bt	common_voice_nl_1		2	0				nl	
21	b51386637c01eace0c2291e27896d7	common_voice_nl_2	eidsvoorste	2	0				nl	
22	bd3c4d0d793087205cd0adeff921bdj	common_voice_nl_2		2	1				nl	
23	bf115c7dcf62b2ac4dcc49baa4d035a	common_voice_nl_2		2	0	fifties	female		nl	
24	bf9594e371b2f199760403701a1769	common_voice_nl_1	Ze hadden naar geveerde deuren ook gesmeerd.	2	1				nl	
25	c18bab9891bf6f1d34a29d6e0dbdbb	common_voice_nl_17700297.mp3	Hij ligt met zijn armen gekruist op de zetel.	2	0				nl	
26	c80d131260865d629ef044c0f4a25bc	common_voice_nl_29432749.mp3	Ook de afgelopen dagen verloren weer tientallen mensen het	2	0				nl	
27	d3441ac81d7899453a061a53222cf1	common_voice_nl_18294089.mp3	De bomen worden eerst gekapt, worden dan in de zagerij tot p	2	1				nl	
28	dcdb0f48072b0c136ddc02a09429e1	common_voice_nl_22499172.mp3	Het is dus een minimumconsensus.	2	0				nl	

How is validation performed?  
Does it match the intended use?  
Is it fair or could it lead to biases?  
Should I provide an alternative validation?

# Looking into the data set metadata

	A	B	C	D	E	F	G	H	I	J
1	client_id	path	sentence	up_votes	down_votes	age	gender	accents	locale	segment
2	0d709133bf209da7f0164653b6e5f9e	common_voice_nl_17699535.mp3	De Aboriginals zijn de oorspronkelijke bewoners van Australië.	2	0	fifties	male	Nederland	nl	
3	0fca93407be6d482019f2463e60fbaf	common_voice_nl_17694848.mp3	Mijn toetsenbord zit vol stof.	2	1			nl		
4	175d4117110538cc68a8a0157a7f0a1	common_voice_nl_18441136.mp3	Ze had de bank beschadigd met haar skateboard.	2	0			nl		
5	29ecd9ca5970fd1bc1297bb4a9a5f3e	common_voice_nl_30205425.mp3	Deze top heeft niet in alle opzichten aan de verwachtingen voldaan.	2	0			nl		
6	3c448c76f72ee15009da6c9e474dc9f	common_voice_nl_19966492.mp3	Waar laat jij je onderhoud doen?	2	0	thirties	male	nl		
7	452cfa9d7e49e72eac2bb7ff3d5f79c	common_voice_nl_17700547.mp3	Na het lezen van vele beoordelingen had ze eindelijk haar oog gevonden.	2	1	fifties	male	Nederland	nl	
8	4ac8a9187cfd11da401d6829199c18	common_voice_nl_21571581.mp3	De tampons zijn op.	2	0	twenties	male	Belgisch N	nl	
9	4d6898ec762a9590a973dc27579b93	common_voice_nl_25310936.mp3	Proberen wij kooldioxide-emissies te beperken of proberen wij ze te verminderen?	2	0			nl		
10	581251914ca5d3e3eb5de368076aecc	common_voice_nl_21431804.mp3	Marijke kent Olivier nu al meer dan twee jaar.	2	0	twenties	female	Nederland	nl	
11	5f1e2fabc5a5c63d880ce655697814e	common_voice_nl_28458010.mp3	Mijn partij vindt dat onverkwijkt.					nl		
12	723a65c399699beeb3c730e141775f	common_voice_nl_19031796.mp3	Het voeren van brood aan een dode is niet toegestaan.					nl		
13	77fc542b71acf6510c80e9a681c128f	common_voice_nl_30125083.mp3	Juni dit jaar is veel te laat.				male	nl		
14	79d570bd61fb0fc4d363d0a1e044cf	common_voice_nl_17695742.mp3	Parket moet je stofzuigen, tegels moet je poetsen.					nl		
15	7aa3af7141d503c7c449404c3f79393	common_voice_nl_19691757.mp3	In onze buurt kent iedereen elkaar.				male	Nederland	nl	
16	8d9feca2b05d9fe37e9c66e45767cae	common_voice_nl_22759350.mp3	Naar mijn oordeel valt er namelijk niets op.					nl		
17	9062e9bc9c6ee752e6f836d4296ee3	common_voice_nl_30057990.mp3	Daarom wordt voorgesteld dit te doen.					nl		
18	a01a8a30db997795dc07295a000bbc	common_voice_nl_17706390.mp3	Heb je de tafel gedekt?					nl		
19	a64ddf48b5b3b6700b146de59126a0	common_voice_nl_17696892.mp3	Ine lag in de strandstoel.					nl		
20	a9f1622fcaedf8cc1fba8d13c58564b1	common_voice_nl_19442483.mp3	Zeven plus vier is elf.					nl		
21	b51386637c01eace0c2291e27896d7	common_voice_nl_24722260.mp3	Het verslag staat echter bol van de neoliberale beleidsvoorstellen.	2	0			nl		
22	bd3c4d0d793087205cd0adeff921bd1	common_voice_nl_20177910.mp3	Ik denk dat ik gevolgd word.	2	1			nl		
23	bf115c7dcf62b2ac4dcc49baa4d035a	common_voice_nl_24733969.mp3	Wij verzetten ons tegen deze begrotingsposten.	2	0	fifties	female	nl		
24	bf9594e371b2f199760403701a1769	common_voice_nl_17703134.mp3	Ze hadden hun gevel bordeauxrood geschilderd.	2	1			nl		
25	c18bab9891bf6f1d34a29d6e0dbdbb	common_voice_nl_17700297.mp3	Hij ligt met zijn armen gekruist op de zetel.	2	0			nl		
26	c80d131260865d629ef044c0f4a25bc	common_voice_nl_29432749.mp3	Ook de afgelopen dagen verloren weer tientallen mensen het leven.	2	0			nl		
27	d3441ac81d7899453a061a53222cf1	common_voice_nl_18294089.mp3	De bomen worden eerst gekapt, worden dan in de zagerij tot splinters vermalen.	2	1			nl		
28	dcdb0f48072b0c136ddc02a09429e1	common_voice_nl_22499172.mp3	Het is dus een minimumconsensus.	2	0			nl		

What accents are present?  
Are foreigners and vulnerable groups represented?

# Looking into the data set metadata

	A	B	C	D	E	F	G	H	I	J
1	client_id	path	sentence	up_votes	down_votes	age	gender	accents	locale	segment
2	0d709133bf209da7f0164653b6e5f9e	common_voice_nl_17699535.mp3	De Aboriginals zijn de oorspronkelijke bewoners van Australië.	2	0	fifties	male	Nederland	nl	
3	0fca93407be6d482019f2463e60fbaf	common_voice_nl_17694848.mp3	Mijn toetsenbord zit vol stof.	2	1				nl	
4	175d4117110538cc68a8a0157a7f0a1	common_voice_nl_18441136.mp3	Ze had de bank beschadigd met haar skateboard.	2	0				nl	
5	29ecd9ca5970fd1bc1297bb4a9a5f3e	common_voice_nl_30205425.mp3	Deze top heeft niet in alle opzichten aan de verwachtingen voldaan.	2	0				nl	
6	3c448c76f72ee15009da6c9e474dc9f	common_voice_nl_19966492.mp3	Waar laat jij je onderhoud doen?	2	0	thirties	male		nl	
7	452cfa9d7e49e72eac2bb7ff3d5f79c	common_voice_nl_17700547.mp3	Na het lezen van vele beoordelingen had ze eindelijk haar oog gevonden.	2	1	fifties	male	Nederland	nl	
8	4ac8a9187cfd11da401d6829199c18	common_voice_nl_21571581.mp3	De tampons zijn op.	2	0	twenties	male	Belgisch N	nl	
9	4d6898ec762a9590a973dc27579b93	common_voice_nl_25310936.mp3	Proberen wij kooldioxide-emissies te beperken of proberen wij ze te verminderen?	2	0				nl	
10	581251914ca5d3e3eb5de368076aecc	common_voice_nl_21431804.mp3	Marijke is een leuke vrouw.	2	0	twenties	female	Nederland	nl	
11	5f1e2fab5a5c63d880ce655697814e	common_voice_nl_28458010.mp3	Mijn partner is een leuke man.	2	0				nl	
12	723a65c399699beeb3c730e141775f	common_voice_nl_19031796.mp3	Het voertuig is een leuke auto.	2	0				nl	
13	77fc542b71ac6510c80e9a681c128f	common_voice_nl_30125083.mp3	Juni dit jaar is een leuke maand.	2	0	seventies	male		nl	
14	79d570bd61fb0fc4d363d0a1e044cf	common_voice_nl_17695742.mp3	Parket met een leuke kleur.	2	1				nl	
15	7aa3af7141d503c7c449404c3f79393	common_voice_nl_19691757.mp3	In onze klas is een leuke leraar.	2	1	teens	male	Nederland	nl	
16	8d9fceb2b05d9fe37e9c66e45767cae	common_voice_nl_22759350.mp3	Naar mijn mening is dit een leuke film.	2	0				nl	
17	9062e9bc9c6ee752e6f836d4296ee3	common_voice_nl_30057990.mp3	Daarom is dit een leuke plek.	2	0				nl	
18	a01a8a30db997795dc07295a000bbc	common_voice_nl_17706390.mp3	Heb je de leukste plek gevonden?	2	1				nl	
19	a64ddf48b5b3b6700b146de59126a0	common_voice_nl_17696892.mp3	Ine lag in een leuke positie.	2	1				nl	
20	a9f1622fcaedf8cc1fba8d13c58564b1	common_voice_nl_19442483.mp3	Zeven plekken zijn er in Nederland.	2	0				nl	
21	b51386637c01eace0c2291e27896d7	common_voice_nl_24722260.mp3	Het verslag staat echter ook voor van de neoliberale beleidsvoorste	2	0				nl	
22	bd3c4d0d793087205cd0adef921bd1	common_voice_nl_20177910.mp3	Ik denk dat ik gevolgd word.	2	1				nl	
23	bf115c7dcf62b2ac4dcc49baa4d035a	common_voice_nl_24733969.mp3	Wij verzetten ons tegen deze begrotingsposten.	2	0	fifties	female		nl	
24	bf9594e371b2f199760403701a1769	common_voice_nl_17703134.mp3	Ze hadden hun gevel bordeauxrood geschilderd.	2	1				nl	
25	c18bab9891bf6f1d34a29d6e0dbdbb	common_voice_nl_17700297.mp3	Hij ligt met zijn armen gekruist op de zetel.	2	0				nl	
26	c80d131260865d629ef044c0f4a25bc	common_voice_nl_29432749.mp3	Ook de afgelopen dagen verloren weer tientallen mensen het leven.	2	0				nl	
27	d3441ac81d7899453a061a53222cf1	common_voice_nl_18294089.mp3	De bomen worden eerst gekapt, worden dan in de zagerij tot planken	2	1				nl	
28	dcdb0f48072b0c136ddc02a09429e1	common_voice_nl_22499172.mp3	Het is dus een minimumconsensus.	2	0				nl	

Are different age and gender types equally represented?  
Could an imbalance lead to discrimination?

# Existing techniques (pros & cons)

- Select a statistically significant sample and check it manually
- Over- or under-sampling
- Make your *testing* data set balanced
- Use automated techniques, e.g., automated gender recognition, to fill the gaps in metadata
- Use multiple data sets and combine them
- Collect new data to integrate missing data
- Use data augmentation and/or synthetic data
- Pre-process your data to add noise or other relevant elements
- Compare your model output with that of different models in different settings

## Conclusion

Pay attention	Pay attention to the source of data: is it reliable?
Assess the data	Assess your dataset in terms of ethics, human rights, and legal requirements
Seek advice	Seek advice from domain experts (doctors, lawyers, statisticians, experts on specific techniques)
Improve the data	Check if you can improve your dataset (ask help from colleagues & doctors in different countries)
Do not use bad data	If your dataset is not good enough, do not use it: do not torture data

ANY  
QUESTION?

[CHIARA.GALLESE@UNITO.IT](mailto:CHIARA.GALLESE@UNITO.IT)

