



DataCom Project

Marie-Sklodowska Curie Action
Chiara Gallese



Funded by
the European Union



ViWebsite
datacomproject.eu



About the project

DataCom started on the 1st of July 2023 at the University of Turin (Turin, Italy) at the Department of Law, under the supervision of Prof. Ugo Pagallo, and will end on the 30th of June 2025.

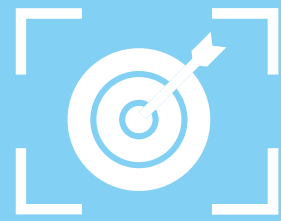
It aims to build a new framework to facilitate an ethical secondary use of health data held by public bodies to improve accountability and enhance responsible reuse.

PI Profile

I am a MSCA postdoctoral fellow working on the intersection of AI and Law, Data Ethics, and Data Protection. My research is highly interdisciplinary and I often collaborate with computer scientists and other scholars from different disciplines.



Objectives & Methodology



Objectives

The project aims at understanding how health data is processed within the Public Administration, what anonymization techniques are employed, and if there are biases in the processes.

It also aims at understanding what are the patients' needs and views regarding the reuse of their health data and if they are open to Data Altruism.

Guidelines will be developed at the end of the project to help public employees in handling the data ethically.



Methodology

I will employ interviews, surveys, focus groups and desk research.

I will administer a survey and interview both patients and public servants in the 3 Member States .

I will also collect relevant sources of law texts and regulations.

I will interview 30 public servants and 30 citizens in each State.

I will also test the guidelines and review them in a focus group with stakeholders.

Why DataCom?

In recent years, health data sets have turned into an economic asset, directly or indirectly monetized by companies and institutions, to such an extent that a health marketplace has been created. They are also being employed in the public sector for many applications: disease control and fighting, pandemic-related decision-making, public health expense estimation, machine learning training, etc.

The conditions under which health data might be legally reused in the EU are various: several laws regulate the reuse of medical data, including GDPR, but the EHDS introduces a novel legal basis that will enhance the possibility of using health data for AI to mitigate biases. This will introduce several research opportunities but also many risks for patients. Detrimental decisions might be taken through AI systems if the data is employed unethically

Image by [DrMo96](#) on Deviantart



How is Health Data reused by the PA?



Performance Evaluation

By understanding the prevalence and distribution of malpractice, dissatisfaction, and claims, governments prioritize funding and services



AI training

Researchers and policymakers use health data to support scientific studies and innovations in healthcare through the use of AI.



Epidemiology

Public administrations analyze health data to conduct epidemiological studies, which provide insights into the causes, patterns, and risk factors of diseases.



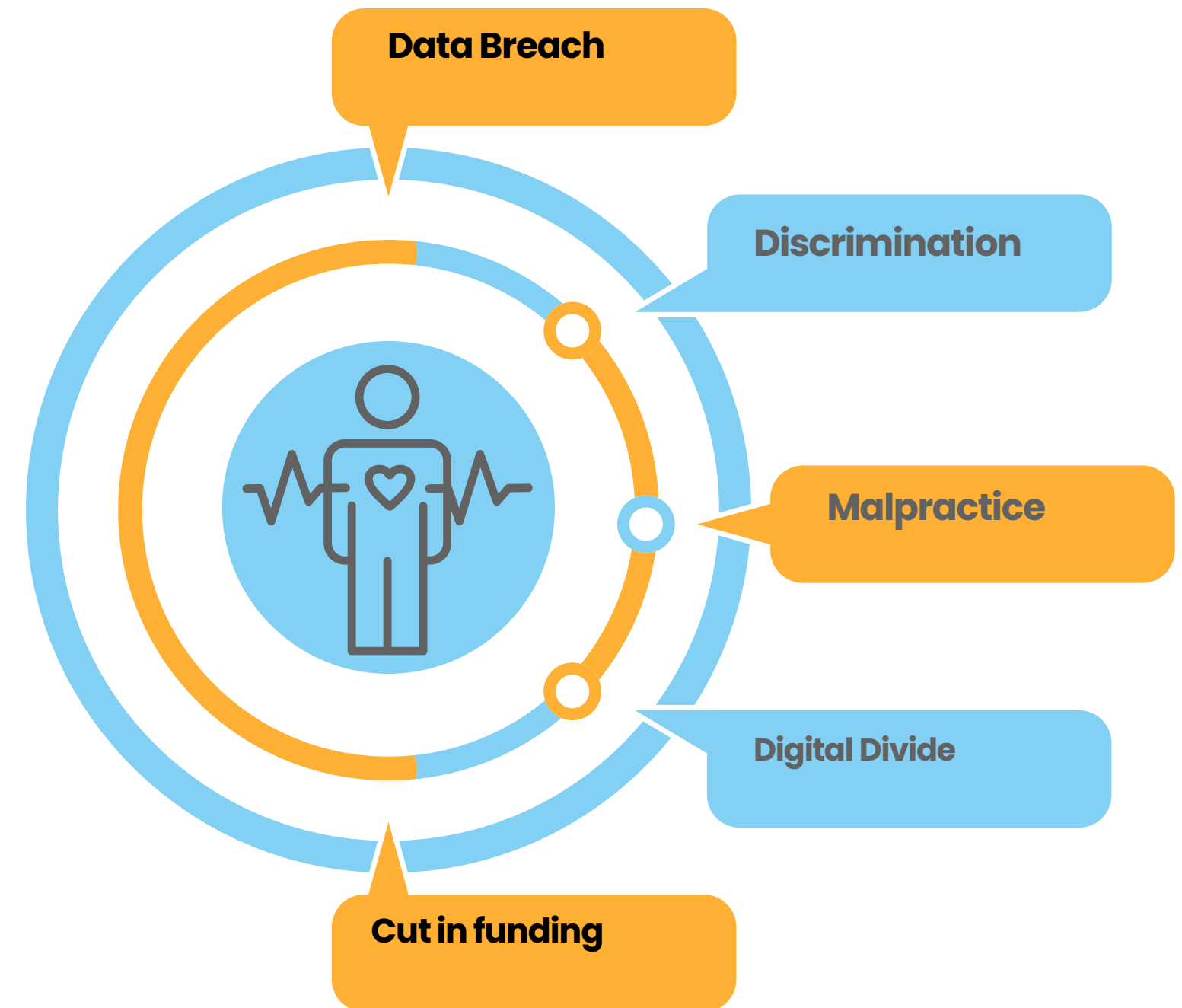
Health Policy

Based on available data, governments plan health policies and make decisions on funding, including resource allocation, planning for healthcare infrastructure, and designing preventive measures.

Risks of reuse and misuse

Although health data and AI can be a great source to improve health in many ways, research has shown that ethical and legal principles should be embedded in the reuse of data to prevent harm to citizens.

Risks associate with use and misuse of health data include re-identification, data breaches, biases and discrimination, medical malpractice, cut in health funding, and limited access to advanced healthcare for some vulnerable and marginalized groups.



BIAS IN HEALTHCARE

Dissecting racial bias in an algorithm used to manage the health of populations

ZIAD OBERMEYER , BRIAN POWERS, CHRISTINE VOGELI, AND SENDHIL MULLAINATHAN  [Authors Info & Affiliations](#)

SCIENCE · 25 Oct 2019 · Vol 366, Issue 6464 · pp. 447-453 · DOI: 10.1126/science.aax2342

↓ 47,479 🗨 1,261



Racial bias in health algorithms

The U.S. health care system uses commercial algorithms to guide health decisions. Obermeyer *et al.* find evidence of racial bias in one widely used algorithm, such that Black patients assigned the same level of risk by the algorithm are sicker than White patients (see the Perspective by Benjamin). The authors estimated that this racial bias reduces the number of Black patients identified for extra care by more than half. Bias occurs because the algorithm uses health costs as a proxy for health needs. Less money is spent on Black patients who have the same level of need, and the algorithm thus falsely concludes that Black patients are healthier than equally sick White patients. Reformulating the algorithm so that it no longer uses costs as a proxy for needs eliminates the racial bias in predicting who needs extra care.





Biases in AI training

In a review published in 2022 on Plos Digital health, it was found that





“U.S. and Chinese datasets and authors were disproportionately **overrepresented** in clinical AI, and almost all of the top 10 databases and author nationalities were from **high income countries** (HICs). AI techniques were most commonly employed for image-rich specialties, and **authors were predominantly male**, with non-clinical backgrounds.”

The problem of human bias







The predominant approach in healthcare involves utilizing **supervised** machine learning models. These models are constructed using historical data that has been **labeled** by medical experts from distinct population segments, acquiring knowledge that is somewhat **limited** in its applicability. The predictions made by these models may exhibit **reduced accuracy** for specific cohorts and could potentially diverge from the suggestions of medical professionals who were not involved in annotating the training data.

A. Clinical data collection

-  1. Examine patients
-  2. Note physiological data
-  3. Fill paper health records
-  4. Transcribe digitally






Systemic bias
Bias in data collection
Bias in transcribing

B. Data set preparation

-  5. Extract data
-  6. Evaluate data
-  7. Preprocess data
-  8. Structure data

Systemic bias within data
Sampling bias
Bias in preprocessing

C. AI creation

-  9. Choose model
-  10. Train model
-  11. Test model
-  12. Refine model
-  13. Deploy AI

Bias in problem formulation
Bias in model design and building
Bias in testing
Bias in AI use

Health Dataset Features for fairness

Compliance

The respect of legal and ethical principles in the whole data life cycle. A data set whose 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.

Quality

The quality of data as assessed by the domain expert (such as if an image is clear or complete, or if the sound can be recognised in a recording).

Completeness

The “holes” in the data set regarding relevant elements, such as missing information about a patient in one or more physiological values.

Numerosity

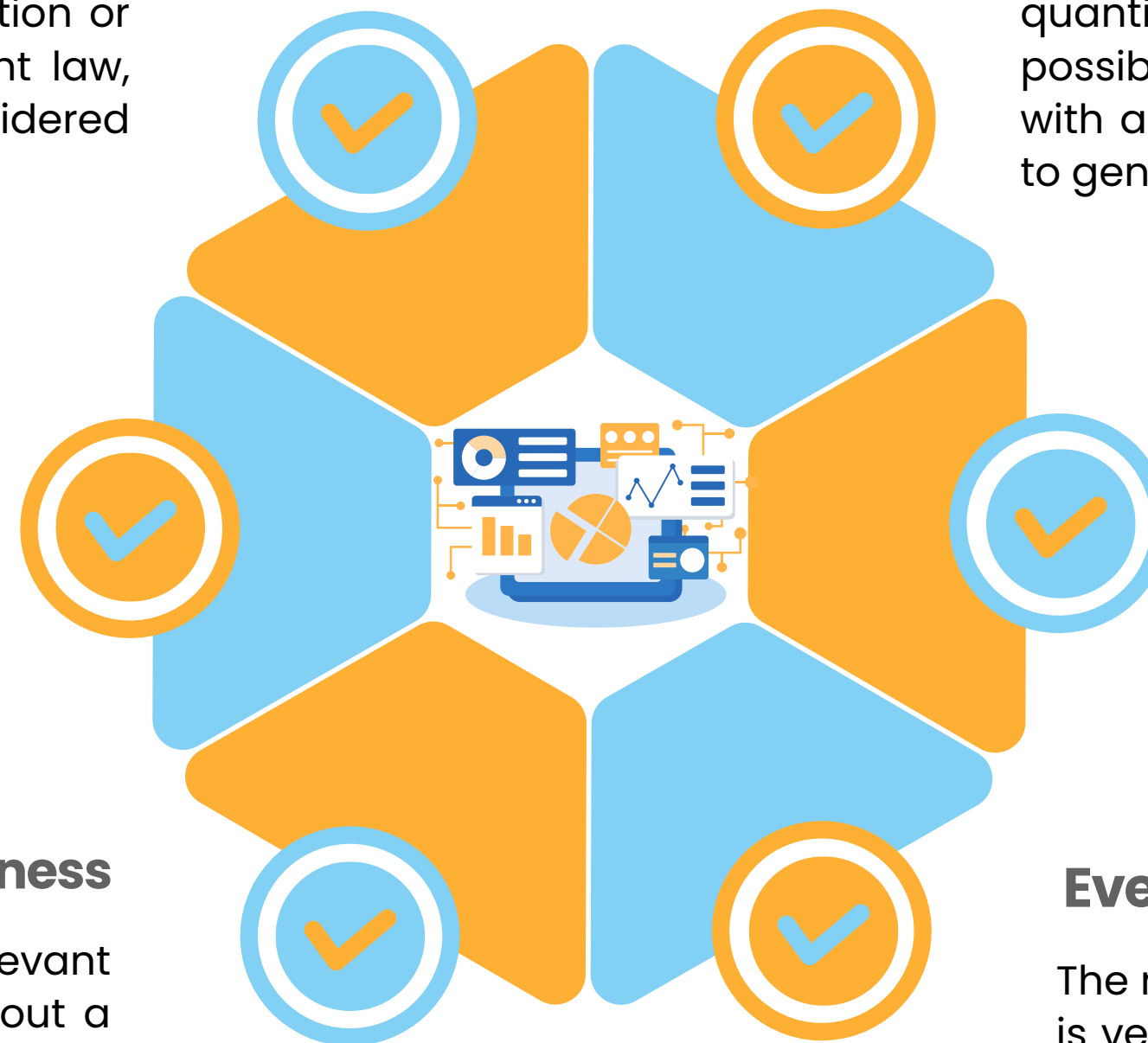
Quantity of instances (S) normalised by the number of features (D) in the data set. The quantity is important in order to mitigate possible biases in the model output: a model with a low number of examples will not be able to generalise.

Balance

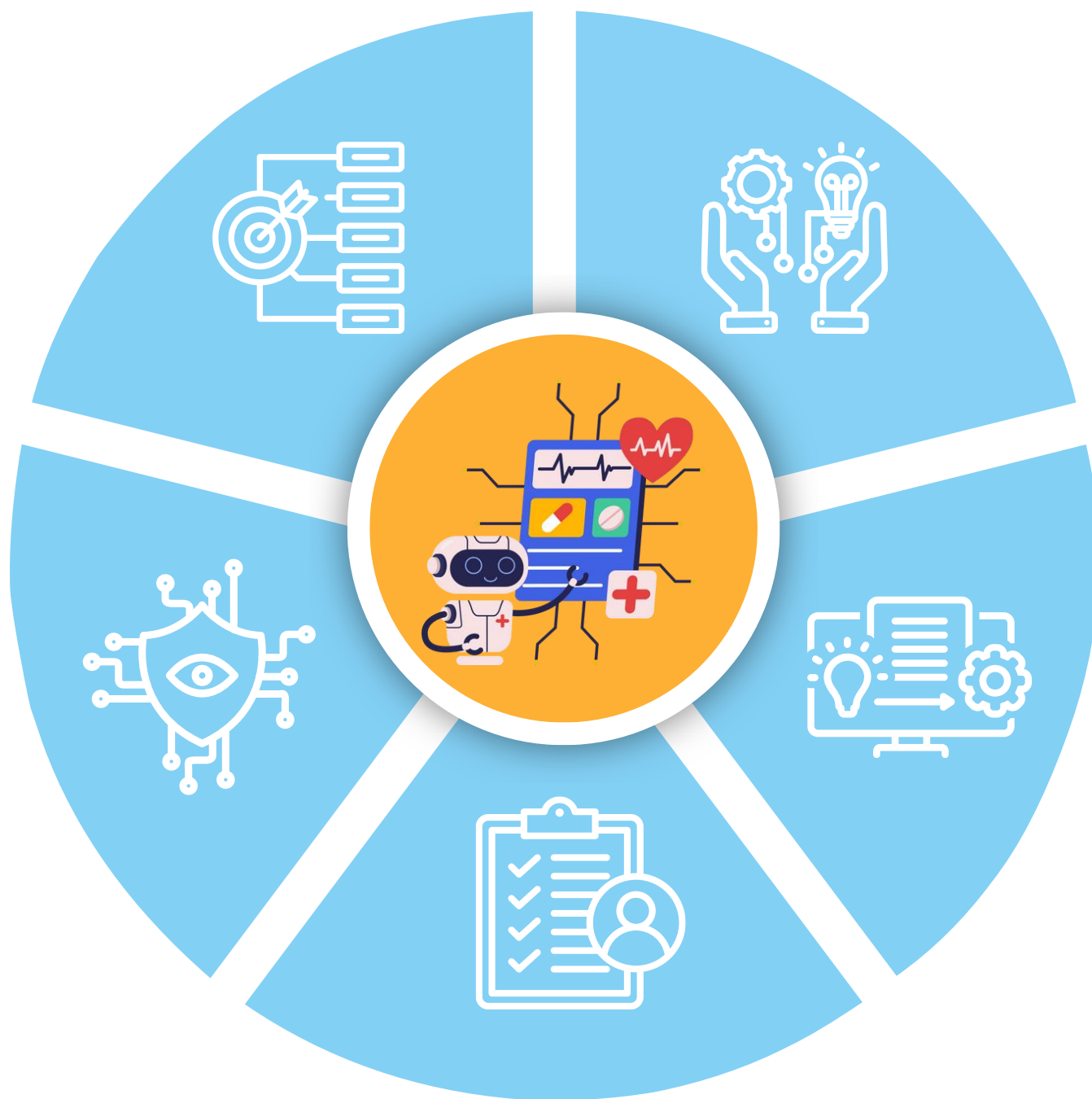
A data set can be defined as “balanced” if all classes are well represented in the data set. An unbalanced data set may produce biased models, unable to generalise and with prejudicial effects on single individuals or entire categories of individuals, such as disparity in accuracy.

Evenness

The number of outliers in the examples. If a value is very different from all other values in the data set, it is important to understand why this occurs.



Corrective measures



Organizational measures

Traning, awarness, bias evaluation, ethical and legal principles implementation



Technical measures

Debiasing, fairness assessment, data preprocessing



Expert consultation

Domain expert inclusion in the AI life cycle



Supervision Or Monitoring

Human oversight, DPIA, fundamental right assessment, risk evaluation



Continuous evaluation

Pre and post-market evaluation, periodic assesment

AI Act Requirements and Obligations

Risk management



Quality Management



Data Governance



Fundamental Right Assessment



Transparency & Information





Anticipatory Regulation: Regulatory Sandboxes and Testing

Conducting pilots to test new regulations in cutting-edge AI application is crucial to understand the issues in practical implementation

Research pilots

Conduct forward-thinking projects to apply new regulations in a controlled testbed



Practical implementation of Law and Ethics

Develop new legal and ethical frameworks to understand how to practically apply relevant principles



Technology co-creation

Create cutting-edge research together with experts and stakeholders, considering citizens' opinions



I am submitting two pilots conducted with Eindhoven University of Technology and LIUC University at the WCCI conference



Thank You

For Your Attention



Visit the project's Website

<https://www.datacomproject.eu>



Contact

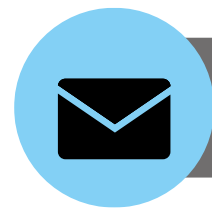
If you want to know more on my research or collaborate in projects, please send me an e-mail or add me on LinkedIn. I will be happy to hear your thoughts on AI & Law!



www.aiandlaw.eu



Chiara Gallese



chiara.gallese@unito.it



Department of Law, Turin University

