30/05/2023

11:30 CET

Reliable and Explainable AI in Trieste

Chiara Gallese, Phd – MSCA PF Fellow 2023
University Of Trieste
Eindhoven University Of Technology
Carlo Cattaneo University LIUC

ITAL – IA - Workshop AI Responsabile e Affidabile

- Monitoring and Verification of Stochastic Systems;
- Embeddings of Logical Formulae; using Graph Neural Networks;
- Formal Methods for Explainable AI;
- Adversarial Robustness;
- Right to interpretability;
- Ethical assessment of data sets;
- AI auditing

Reliable and Explainable AI in Trieste

Emanuele Ballarin[1], Luca Bortolussi[1], Francesca Cairoli[1], Chiara Gallese[1], Laura Nenzi[1] and Gaia Saveri[1,2]

[1]*Department of Mathematics and Geoscience, University of Trieste, Italy*

[2]*Department of Computer Science, University of Pisa, Italy*

Abstract

This paper summarizes the activity in the area of Reliable and Explainable AI carried out at the University of Trieste. The main topics are: monitoring and verification of stochastic systems, embeddings of logical formulae using Graph Neural Networks, formal methods for explainable AI, adversarial robustness, right to interpretability, ethical assessment of data sets and AI auditing.

Table of contents

01 The concepts of “Technical Interpretability” and “Explainability”

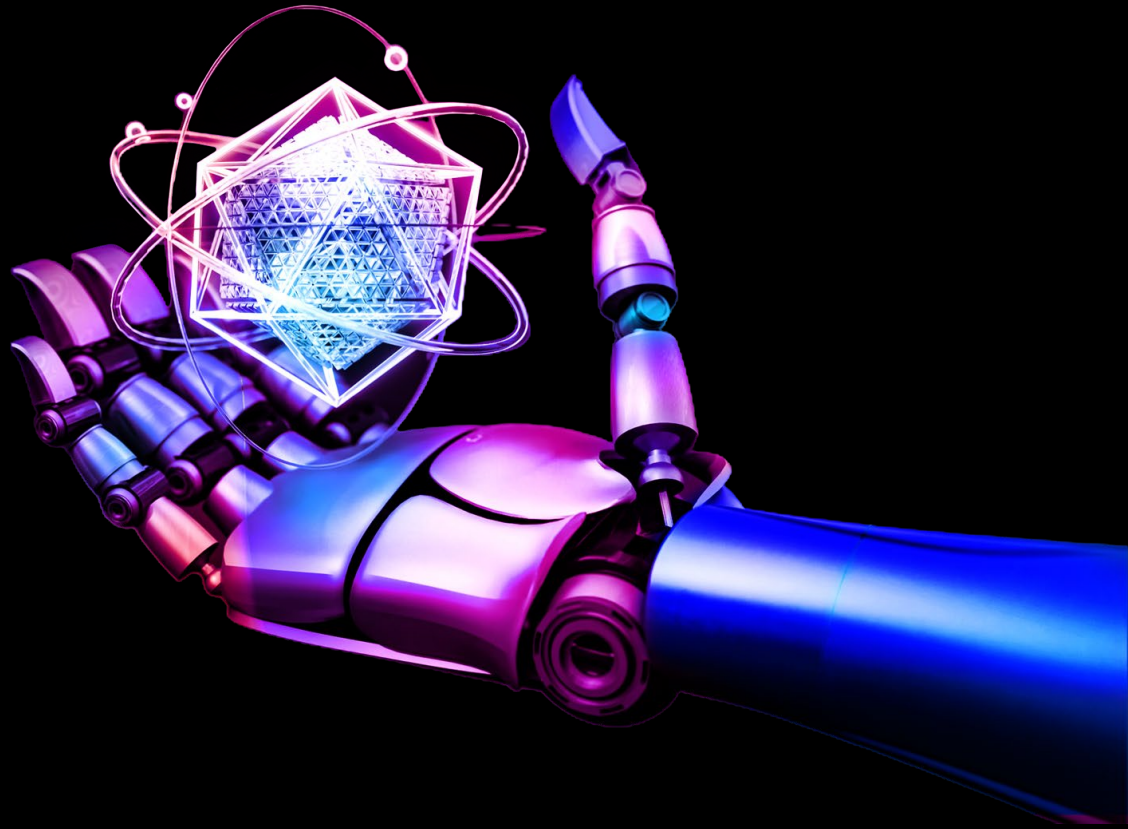
02 The transparency principle and the “Right to Explanation”

03 Articles 13 and 14 of the AI Act Proposal

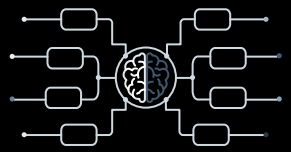
04 The “Right to Technical Interpretability” as a fundamental right

01

The concepts of
“Technical Interpretability”
and
“Explainability”

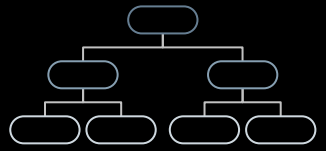


definitions



Explainability

A post-hoc model built to understand how the original black box model reached a certain conclusion



interpretability

A white box model in which it is clear what the underlying reasoning is

EXAMPLES

Consider a classification algorithm that categorizes upcoming patients of a particular condition into healthy or unhealthy based on a data set comprising millions of physiological characteristics of past patients. In order to determine whether the system classifies a new patient as healthy or ill, doctors may enter data from the patient, including blood levels, symptoms, anamnesis, genomic information, lifestyle choices, age, number of children, ethnicity, weight, height, number of sleep hours, job, place of birth, etc. Because of the numerous and intricate features, parameters, and layers that are employed in producing the output, the system is unable to determine the cause of the patient's illness, such as the fact that the blood levels are abnormal for someone of the patient's age, ethnicity, weight, and exercise.

Considering a system that predicts the likelihood of not being able to pay back a mortgage and is used by a financial institution to deny credit, we would consider it interpretable only if it made clear which financially significant factors—such as wage, job type, age, concurrent loans, marital status, and education—were used by the model to produce the output, what relationship were found between them (e.g., educated persons are more likely to have high incomes), and which ones were given a higher weight than others (for example, the system could weight the past mobility as an unfavourable condition and weight it more that an advanced age.

02

The transparency principle and the "Right to Explanation"

transparency

Transparency is a key principle and an overarching obligation in the whole EU legislation and within the Digital Strategy, but it is also an important ethical and legal requirement provided by national laws and guidelines in some fields relating to high-risk systems.

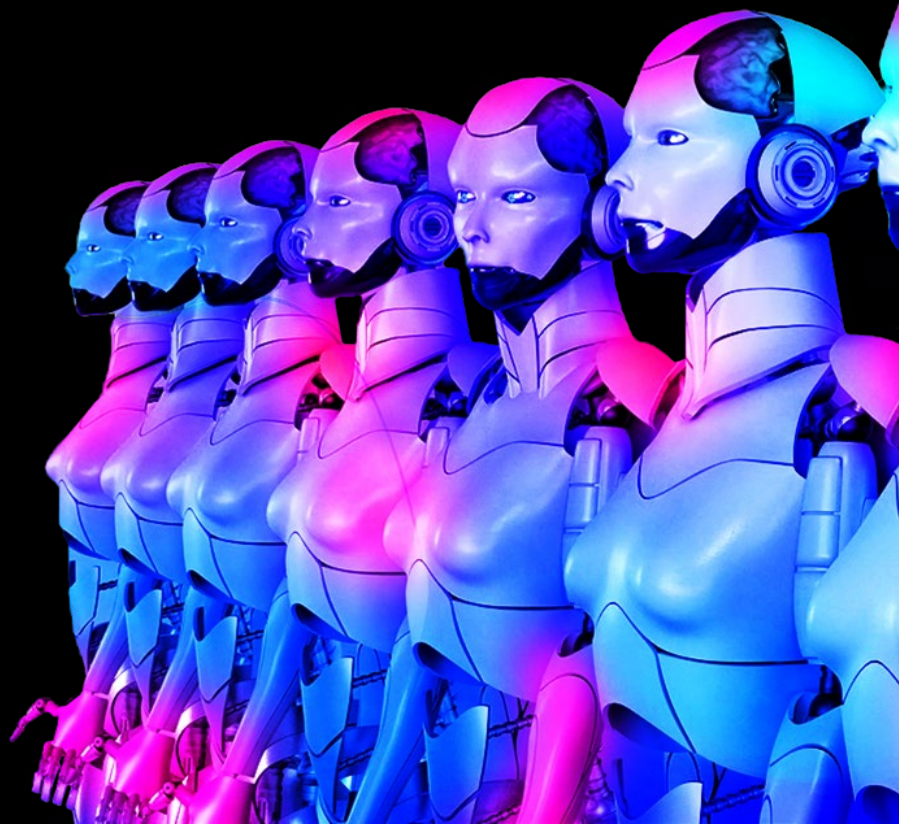
The "right of explanation" in GDPR is part of transparency: data subjects have the right to receive information about the rationale behind or the criteria relied on in reaching an automated decision that has an impact on their life, and about the significance and envisaged consequences of the processing of their data, as provided by Articles 13 and 14 of GDPR.

Convention 108+, article 10

"Data subjects should be entitled to know the reasoning underlying the processing of their data, including the consequences of such reasoning, which led to any resulting conclusions, in particular in cases involving the use of algorithms for automated decision making including profiling. For instance, in the case of credit scoring, they should be entitled to know the logic underpinning the processing of their data and resulting in a 'yes' or 'no' decision, and not simply information on the decision itself. Without an understanding of these elements, there could be no effective exercise of other essential safeguards such as the right to object and the right to complain to a competent authority".

03

Articles 13 and 14 of the AI Act Proposal



“

Transparency and provision of information to users 1. High-risk AI systems shall be designed and developed in such a way to ensure that their operation is sufficiently transparent to enable users to interpret the system’s output and use it appropriately. An appropriate type and degree of transparency shall be ensured [..] 3. The information referred to in paragraph 2 shall specify: [..] (d) the human oversight measures referred to in Article 14, including the technical measures put in place to facilitate the interpretation of the outputs of AI systems by the users [..]”

—AI ACT, ARTICLE 13

Article 14

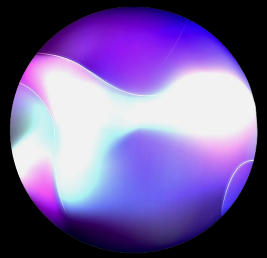
Article 14 mentions the concept of interpretability when referring to the human oversight measures, prescribing that one of the measures to achieve it is to enable the user to “correctly interpret the high-risk AI system’s output, taking into account in particular the characteristics of the system and the interpretation tools and methods available”.

Interpretability is then a mandatory, yet alternative, measure to make sure that a human is always kept in the loop to oversee the behavior of the AI system.

04

The "Right to Technical Interpretability" as a fundamental right

The right to technical interpretability

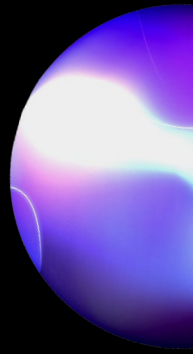


Bias detection

In black boxes, it is not possible to detect biases in advance

Bias mitigation

In black boxes, it is not possible to mitigate biases



Informed consent

Only white boxes guarantee a real informed consent

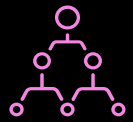
Trust

White boxes generate more trust

Right to challenge

You cannot challenge what you don't know

solutions



Multimodal ai

Build combined models which guarantee interpretability of decisions

Black boxes

Use black boxes only when really needed (high accuracy, image analysis)

Human oversight

Enact non-technical measures to guarantee an effective human intervention in the decision

Thanks!

Do you have any question?

chiara.gallese@units.it