Explainable artificial intelligence

EDPS

EUROPEAN DATA PROTECTION SUPERVISOR

The EU's independent data protection authority

Vítor Bernardo

30/11/2023

Why is transparency important in AI?

To prevent the '*black-box*' effect

Difficult to predict discriminatory decisions or behaviours

Difficult for data subjects to understand why/how they can be impacted

Difficult to identify the limitations of the system (e.g. Clever Hans effect)

Common pitfalls of «black-box» systems

The algorithm performed well during tests

Tested under same conditions as where it will be deployed?
Can the controller demonstrate that the algorithm is fair and accurate?

The system does not produce automated decisions (as in Articles 24 and 77 EUDPR)

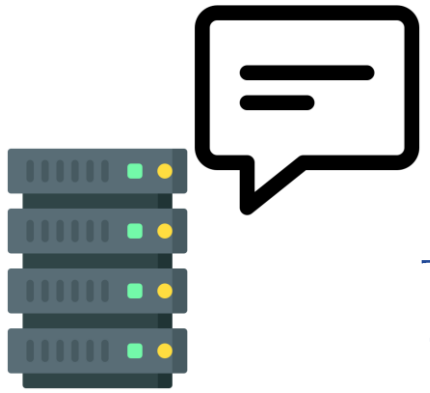
Decision support systems influence human decisions,
especially when confirming already existing biases

Users are sufficiently aware of the risk of using the system

Users may be led to believe that system suggestions are valid
because they are provided by the system.

What is explainable AI (XAI)?

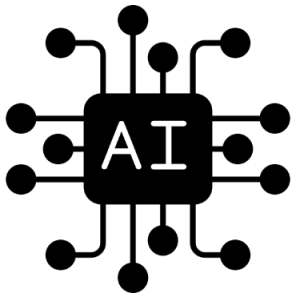
Ability of AI systems to provide clear and understandable explanations for their actions and decisions



The goal is to make the behaviour of the systems understandable to humans by elucidating the underlying mechanisms of their decision-making processes

Possible approaches to explainable AI

Less complex



More complex

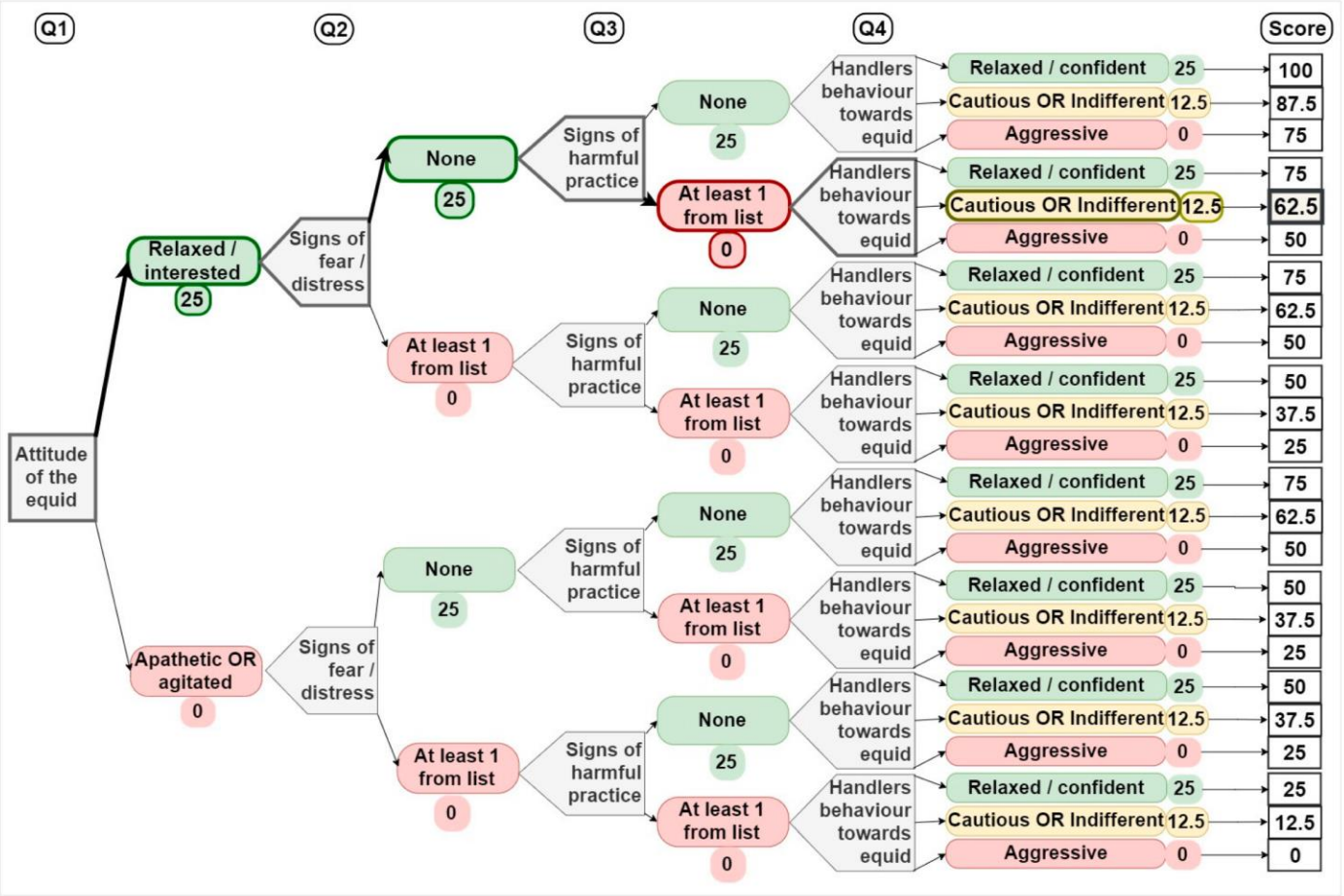
Self-interpretable models

e.g. decision trees, linear regression

Post-hoc explanations

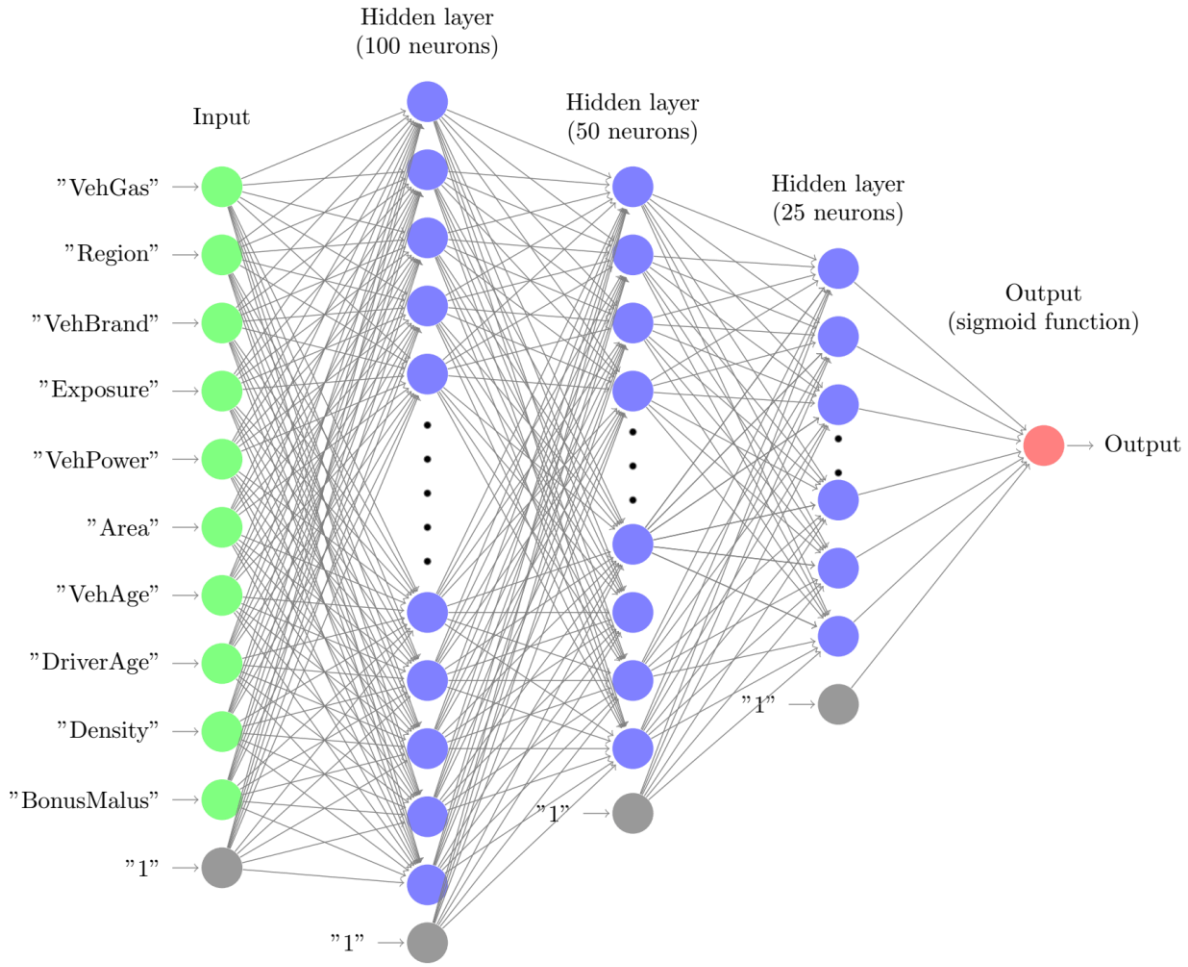
e.g. deep learning

Example of a self-interpretable model (decision tree)



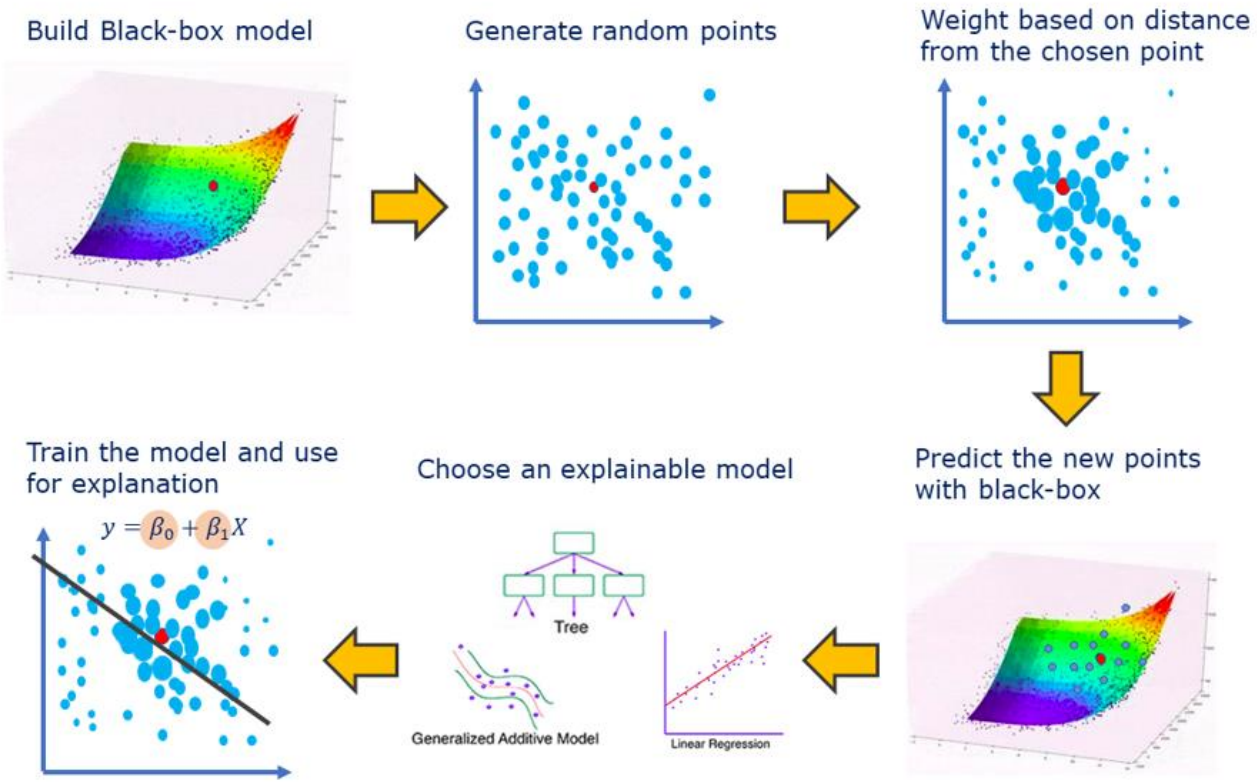
Source: The Welfare Aggregation and Guidance (WAG) Tool: A New Method to Summarize Global Welfare Assessment Data for Equids, https://doi.org/10.3390/ani10040546

Example of a deep learning model



Source: 9 Steps of Debugging Deep Learning Model Training, https://neptune.ai/blog/debugging-deep-learning-model-training

Post hoc explanations: LIME



LIME stands for Local Interpretable Model-agnostic Explanations

Picture of the of the LIME algorithm steps by Giorgio Visani

Post hoc explanations: not a silver bullet

"(...) the use of post-hoc explanatory methods is useful in many cases, but these methods have limitations that prohibit reliance as the sole mechanism to guarantee fairness of model outcomes in high-stakes decision-making".

Vale, D. E.-S. (2022). Explainable artificial intelligence (XAI) post-hoc explainability methods: Risks and limitations in non-discrimination law. AI and Ethics

"(...) from a technical and philosophical point of view these explanations can never reveal the "unique, true reason" why an algorithm came to a certain decision (...) in the worst case, the explanations may induce us into falsely believing that "justified", or "objective" decision has been made even when this is not the case".

Bordt, S. F. (2022). Post-hoc explanations fail to achieve their purpose in adversarial contexts. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency

Therefore, the limitations of "black box" approaches should be considered when trying to assess the fairness of the models.

XAI and personal data

Transparency

Accountability

Data minimisation

Special categories of data

Risks associated with the implementation of XAI

Misinterpretation

Over-reliance on the AI system by deployers

Potential exploitation of the systems

Disclosure of trade secrets

The importance of the human factor

Humans prefer
contrastive explanations

Explanations
are social

Humans
are selective

Explanations are
contextual

Humans must
trust explanations

Suggestion:

Questions for assessing transparency

Questions for assessing transparency*

- What types of AI are used? (more/less complex)
- Are the AI models developed in-house?
- Is the code open-source?
- Does the systems provide explanations on its functioning?
- Are the explanations tailored to different audiences? Which ones?

* This set of questions serves as a reference for self-assessment of transparency and is in no way intended to be exhaustive or to ensure compliance. Assessing the transparency of a system requires a case-by-case analysis.

Questions for assessing transparency*

- On the basis of the explanations, is it possible to identify:
  - what categories of data are relevant to the decision?
  - how does the system process the data to arrive at decisions?
  - the limitations of the system?

- Is the data subject able to request a different explanation if the one provided is not considered sufficiently clear?

- Does the system provide mechanisms for auditability (e.g. by a designated supervisory authority)?

* This set of questions serves as a reference for self-assessment of transparency and is in no way intended to be exhaustive or to ensure compliance. Assessing the transparency of a system requires a case-by-case analysis.

Follow more technology monitoring of the EDPS on..

Follow more technology monitoring of the EDPS on..

[TechSonar report 2021-2022]
[TechSonar report 2022-2023]
TechSonar report 2023-2024 (to be published in a few days)

[Explainable Artificial Intelligence (XAI)]
[Synthetic data: "what use cases as a privacy enhancing technology?"]

[TechDispatch #2/2023 - Explainable Artificial Intelligence] (listen the podcast here)
[TechDispatch #1/2021 - Facial Emotion Recognition]

EUROPEAN DATA PROTECTION SUPERVISOR

The EU's independent data protection authority

vitor.bernardo@edps.europa.eu

All icons from https://www.flaticon.com/
All images from https://www.flickr.com, except where indicated

@EU_EDPS

European Data Protection Supervison

EDPS