Convention A:
Algorithm Auditing Ensuring Trustworthy

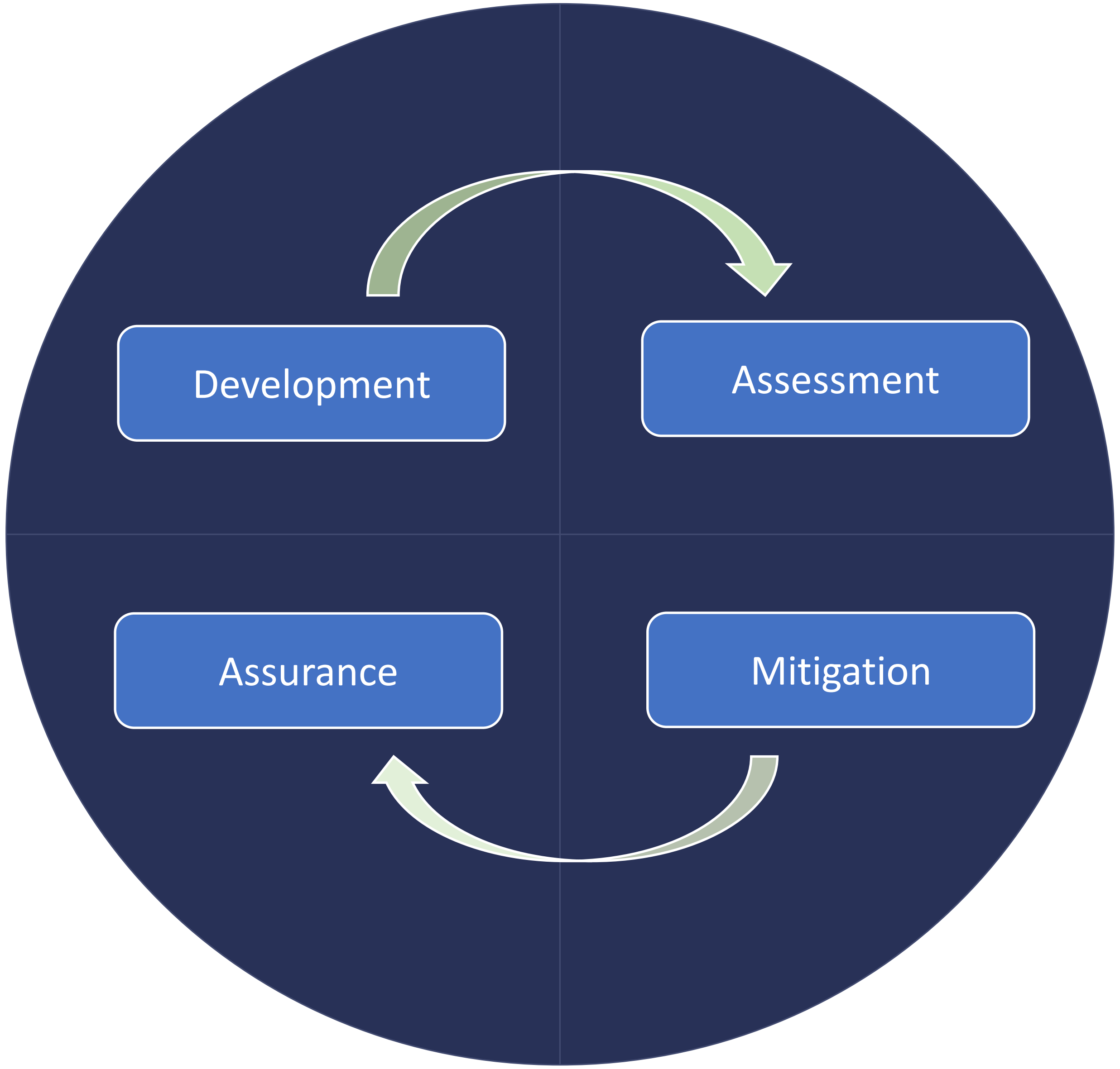
Francesco Toraldo
Claudio Senatore Reso

19 March, 2024

The Audit Process

Developing and documenting an algorithmic system

Evaluating the algorithm behaviour and capacities

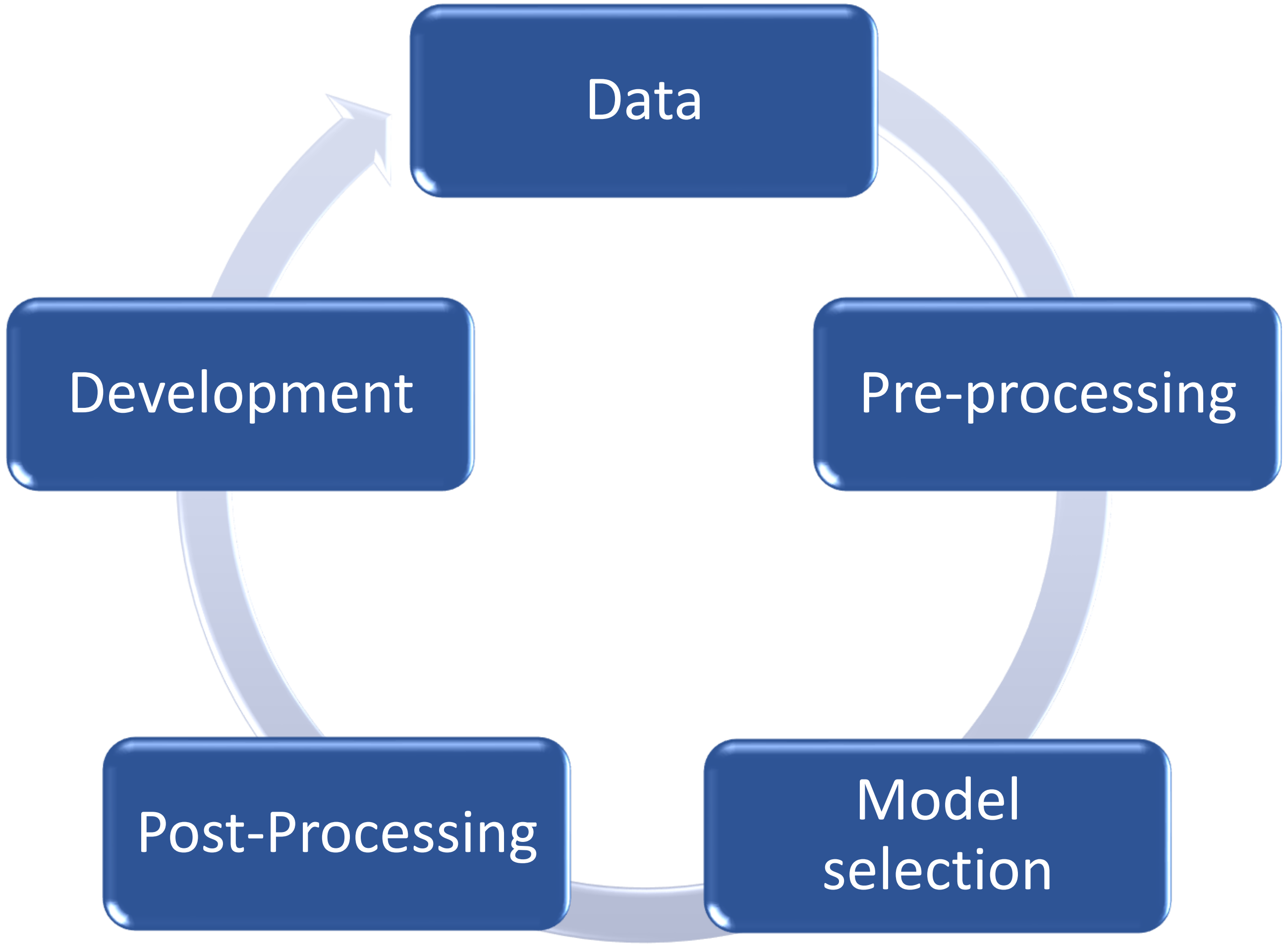


Risk rating, Certification, Best Practice

Servicing or improving an algorithm outcome

The Audit Matrix

The Algorithm Audit section aligns with the general stages of a standard modeling project, yet its internal workings are distinct.



Data

Pre-processing

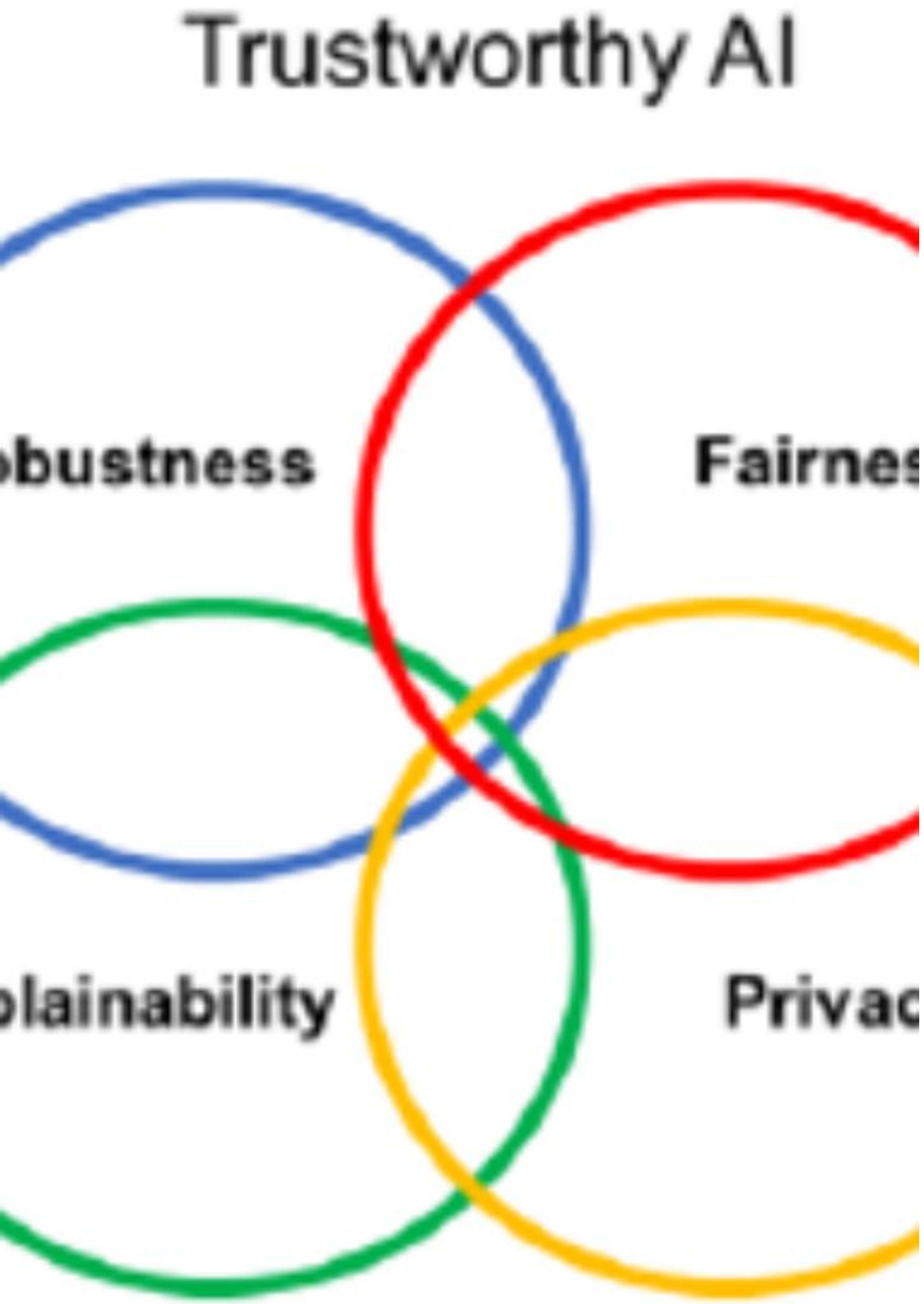
Model selection

Post-Processing

Development

The Audit Matrix: Interrelation between Development Stage and and Auditing Verticals

Stage	Explainability	Robustness	Fairness	Privacy
Data and Task Task Setup	Data collection collection and and labelling	Data Accuracy	Population balance	Data Protection Impact Assessment
Feature pre-processing	Dictionary of variables	Feature engineering	Fair representation	Data minimization
Model Selection	Model complexity	Model validation	Fairness constraints	Differential privacy
Post-Processing and Reporting	Auxiliary tools	Adversarial testing	Bias metrics assessment	Model Inversion
Productionizing and Deploying	Interface and documentation	Detecting Concept Drift & Continuous Integration	Real-time monitoring of bias metrics	Managing User Queries and Implementing Rate-Limiting



Trade-offs and Interactions

The interplay between different auditing verticals often involves trade-offs, where optimizing one aspect may affect another. For instance, enhancing accuracy might come at the cost of fairness, or increasing explainability could impact privacy. These trade-offs are not only technological considerations but also reflect the values and regulatory requirements of the organization. Understanding and managing these trade-offs is crucial for developing trustworthy AI systems.

Interpretability and Explainability

Interpretability

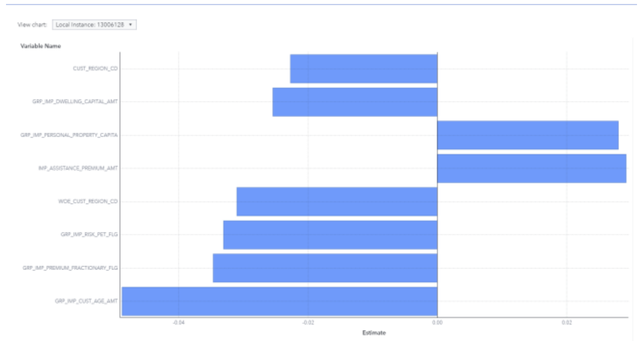
Understanding the cause and effect within the system and predicting outcomes based on inputs or parameters.

Explainability

Providing human-understandable explanations for the internal mechanics of machine learning systems.

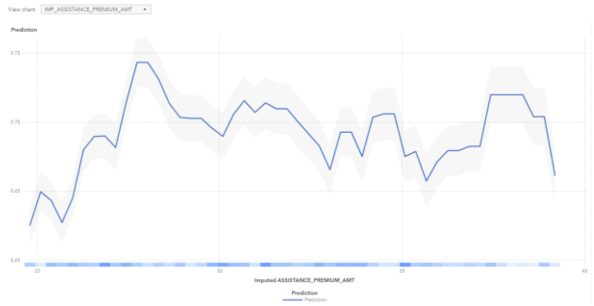
Explainability

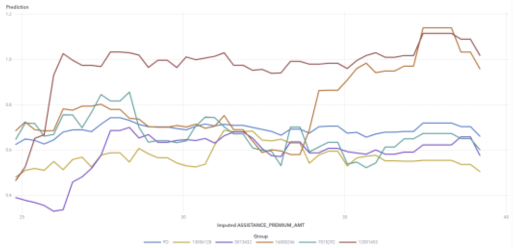
Model-agnostic



- LIME
- Shapley value (SHAP)

- Partial Dependence
- Feature Attribution
- Adversial Perturbations
- UCFE





- ICE

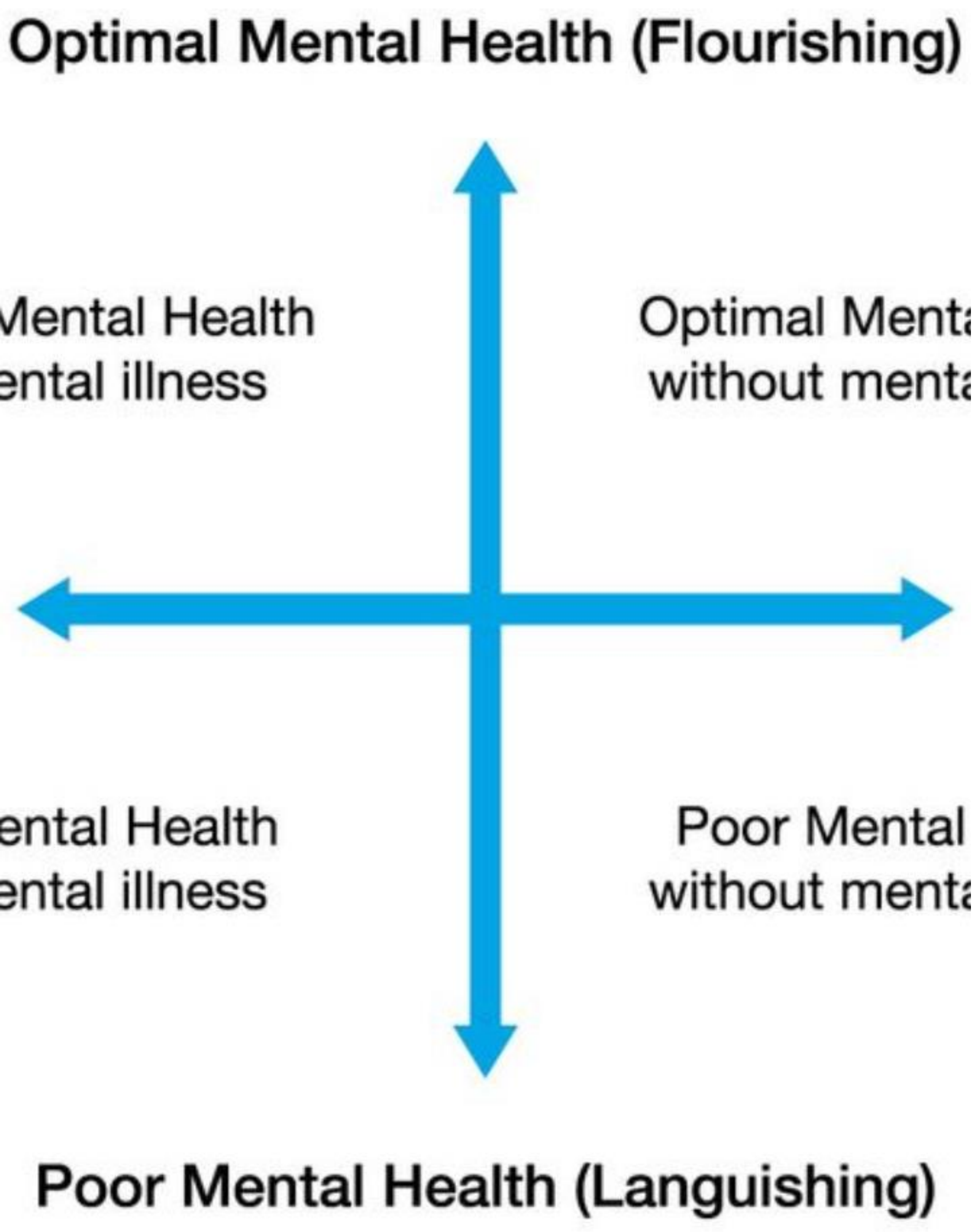
Local

Global

- Linear Model

- Decision Tree
- PROBE
- Attention Mechanisms

Model-specific



Algorithm Auditing: A Continuum

Algorithm auditing is not a binary process but rather a continuum, with varying shades of access and knowledge. From 'White-box' to 'Black-box', each level of access provides different opportunities and challenges for auditors, requiring a nuanced approach to evaluate and improve algorithmic systems effectively.

Bias creeping in to AI

Historical Bias

Perpetuation of historical bias, amplification of amplification of bias

Highly Influential Data Points

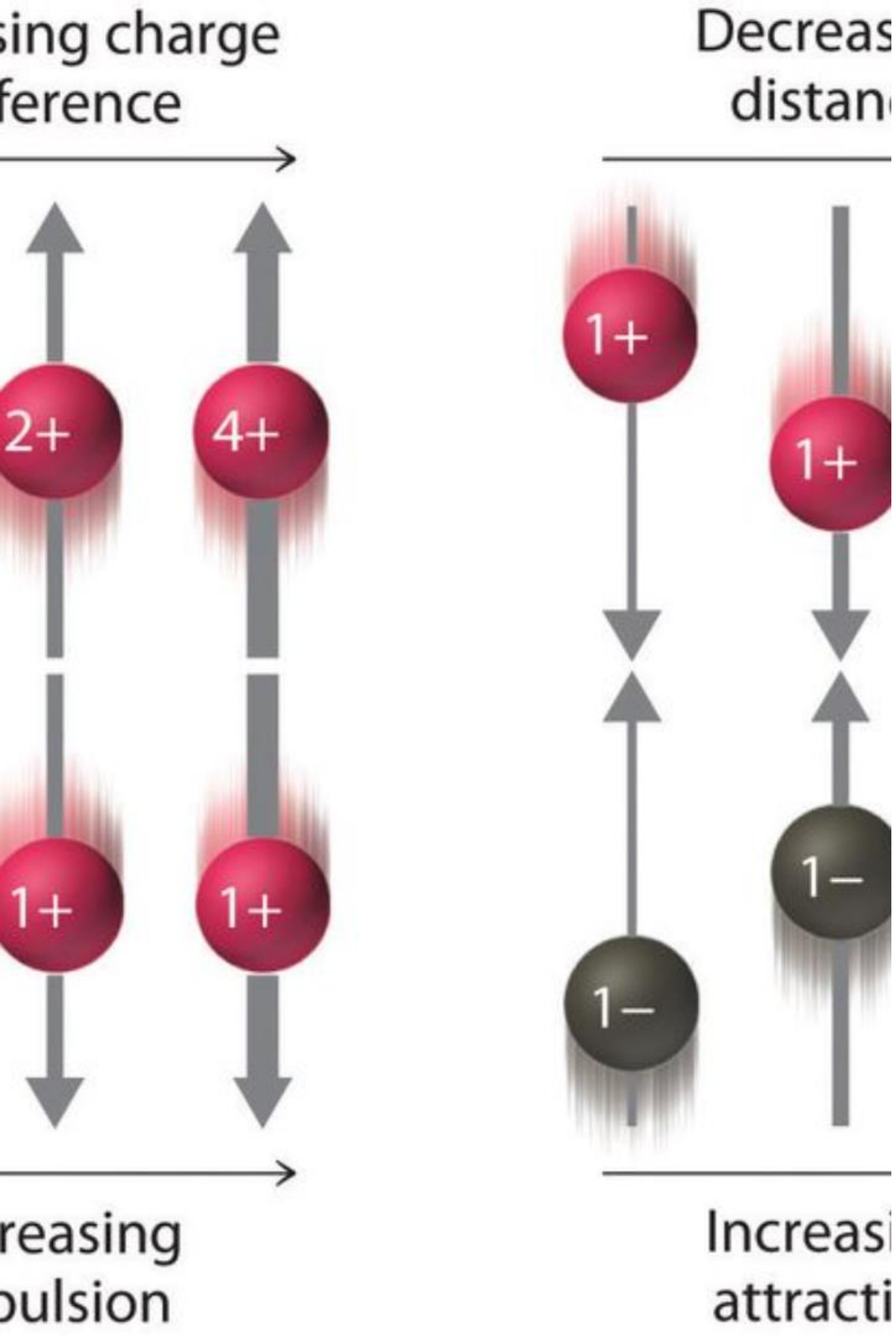
Preventing self-justifying feedback loops that that reinforce existing biases.

With-in model bias

Subsampling (e.g. Random Forest), sampling within within the algorithm can inadvertently subsample subsample poorly, especially for rare events.

Surrogate and Proxy Bias

If remove sensitive variables, won't be able to know if if surrogates/proxies are biasing. Moreover removing a a variable might not be sufficient.



ISOA

Explainability vs Robustness

Striking a balance between making an algorithm understandable and maintaining its performance is a delicate task. Enhancing explainability can sometimes compromise robustness, as simplifying the model for clarity might reduce its predictive power. Conversely, a highly robust algorithm could become a 'black box,' difficult to interpret.

Fairness vs Robustness

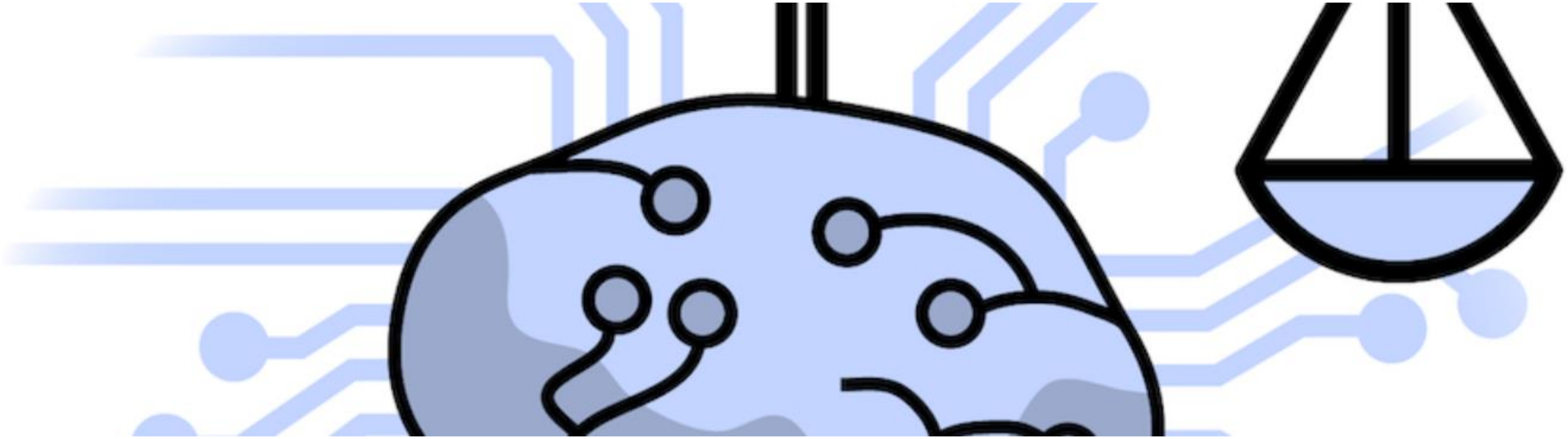
The interplay between fairness and robustness is a critical aspect of algorithm auditing. Ensuring that an algorithm is fair often requires adjustments that may affect its accuracy or performance. For example, incorporating fairness constraints might limit the algorithm's ability to learn complex patterns, potentially reducing its robustness.

ISOA

Explainability vs Privacy

Achieving a harmonious blend of explainability and privacy in algorithmic decision-making poses a significant challenge. Explainability relies on transparency in data processing, while privacy focuses on preserving the confidentiality of sensitive information. Reconciling these objectives entails meticulous consideration of the level of detail disclosed about the algorithm's internal workings and the strategies used to maintain data protection.

ISOA

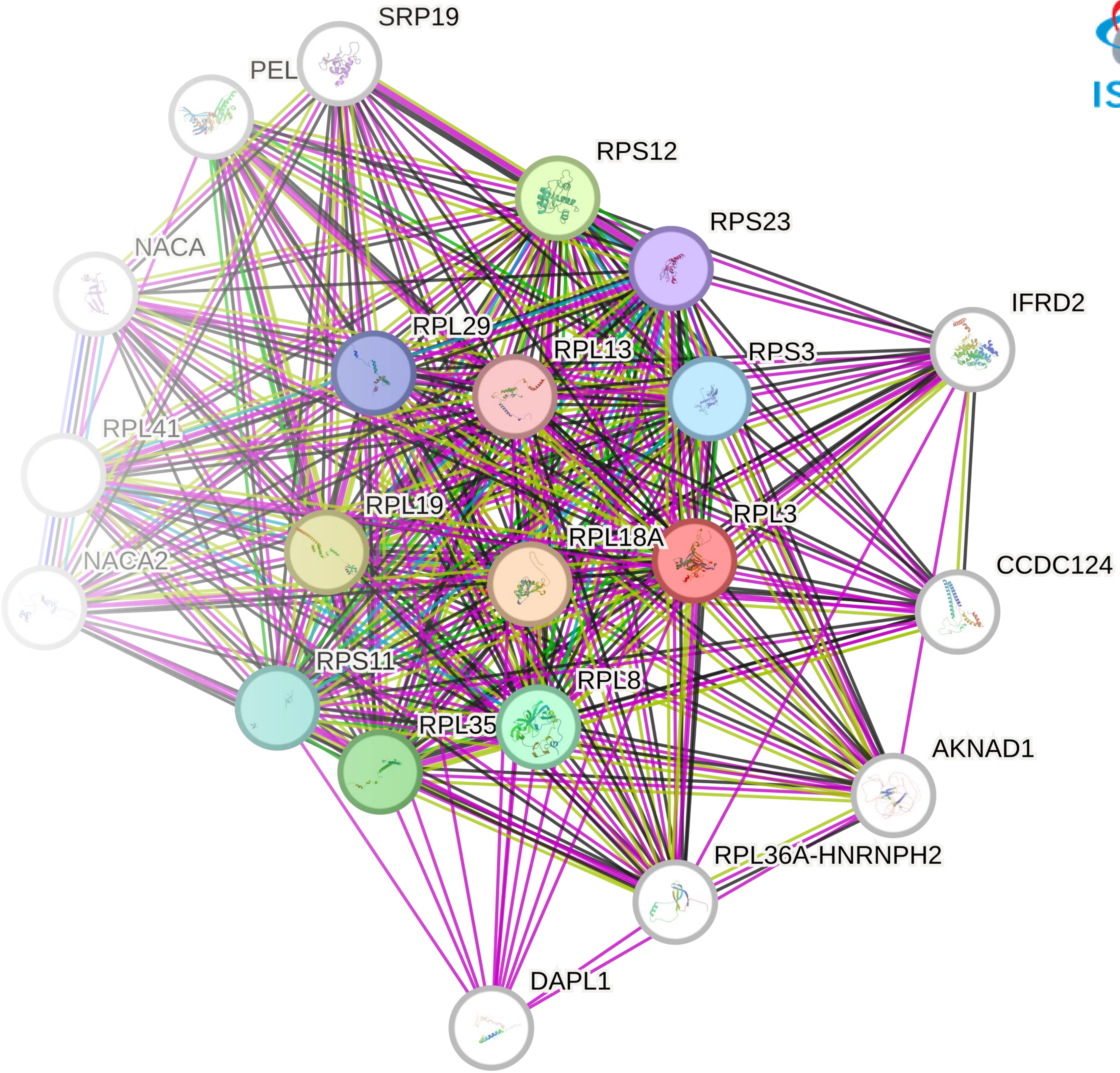


Fairness vs Explainability

Enhancing an algorithm's explainability can contribute to increased transparency, facilitating the identification of potential biases and fostering fairness. Nevertheless, the process of increasing explainability may unintentionally expose sensitive information, which could be misused for discriminatory purposes. Consequently, auditors must strike a delicate balance to guarantee that improvements in explainability do not unintentionally perpetuate or exacerbate discriminatory practices.

Interaction between all Verticals

Various dimensions of algorithm auditing interact closely, forming an intricate web of dependencies and consequences. Improving performance might have implications for fairness or privacy, while enhancing robustness can influence explainability. Gaining insight into these interconnections is vital for holistic auditing and fostering the development of reliable, trustworthy AI systems.

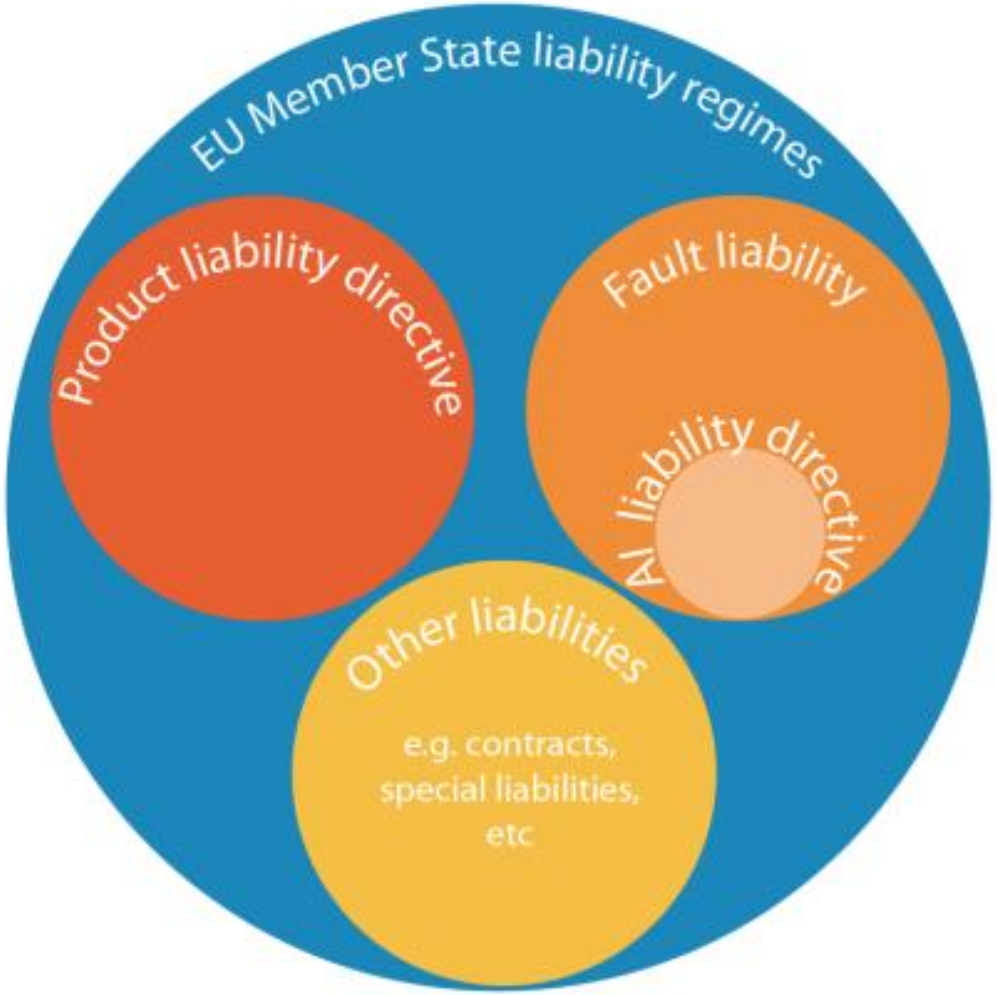


Focus on EU AI liability

PLD: it has been reached interinstitutional agreement on a final text.

AILD: still under discussion. Last update 28/09/22

Art 18 (PLD) and Art 3 (AILD) state that claimant only need to provide plausible evidence of potential harm, while defendants must disclose all relevant information to avoid liability, with non-compliance to this disclosure leading to a (rebuttable) presumption that the defendant has breached its duty of care.


EU Member State liability regimes
Product liability directive
Fault liability
AI liability directive
Other liabilities
e.g. contracts, special liabilities, etc