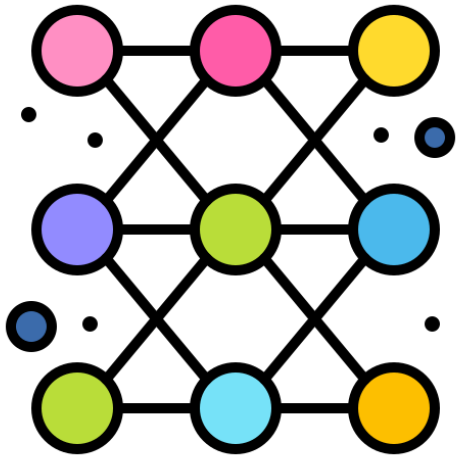
Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI

By:

Saman Attarkashani

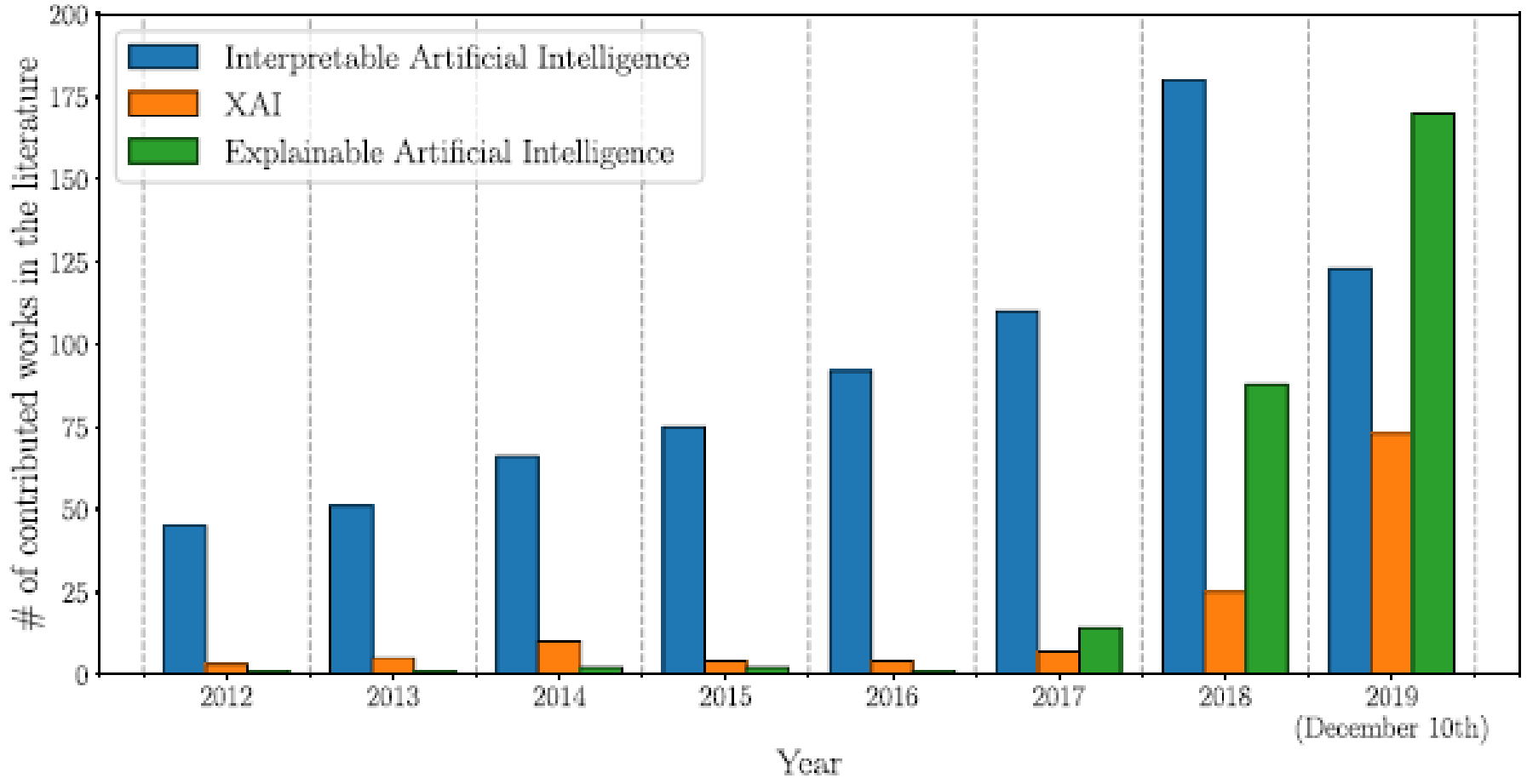
Supervisor:

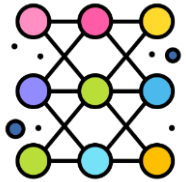
Dr. Azadeh Mansouri



Introduction

Introduction





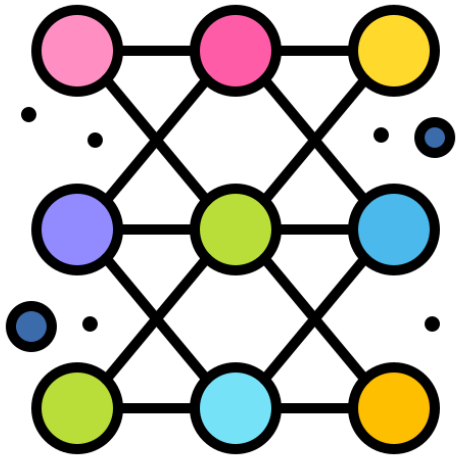
Introduction

When developing a ML model, the consideration of interpretability as an additional design driver can improve its implementability for 3 reasons:

Interpretability helps ensure impartiality in decision-making, i.e. to detect, and consequently, correct from bias in the training dataset.

Interpretability facilitates the provision of robustness by highlighting potential adversarial perturbations that could change the prediction.

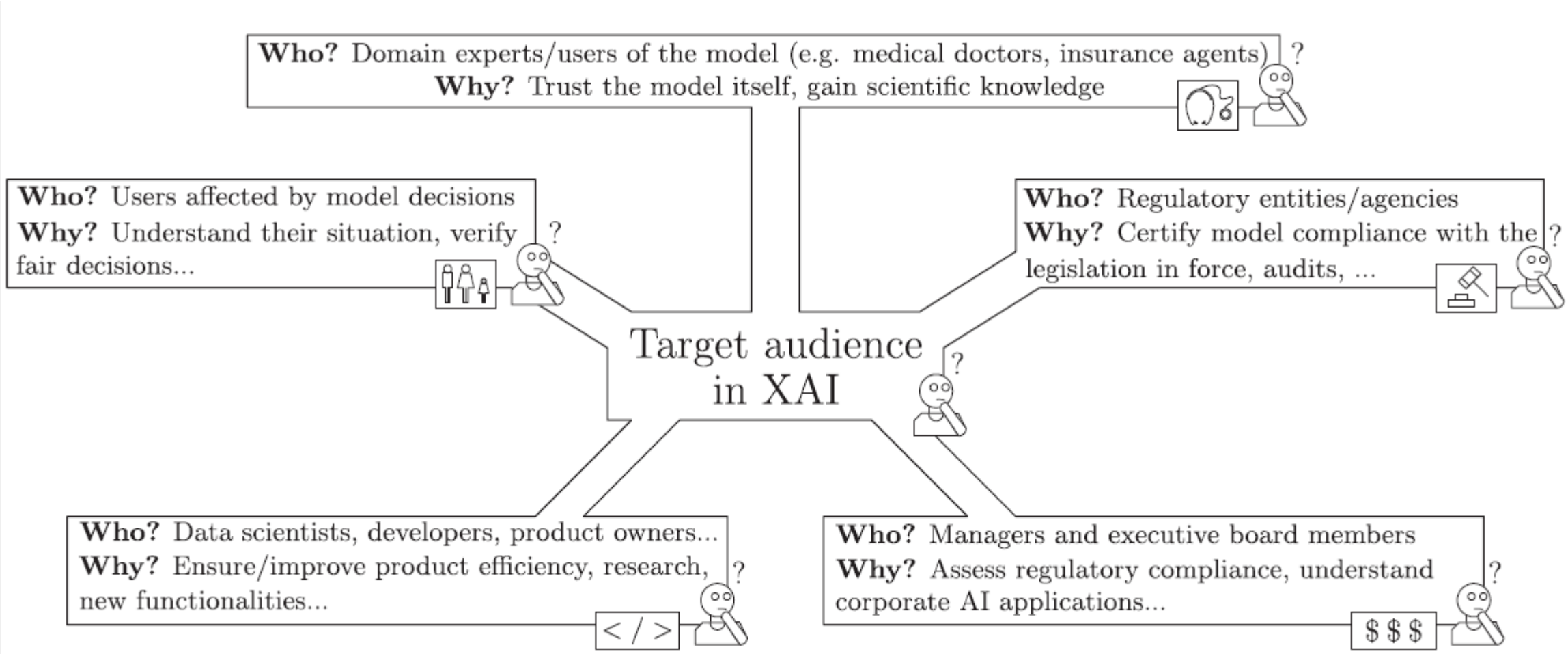
Interpretability can act as an insurance that only meaningful variables infer the output, i.e., guaranteeing that an underlying truthful causality exists in the model reasoning.

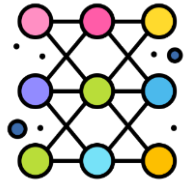


Explainability: What, why, what for and how?

What, why, what for and how?

Diagram showing the different purposes of explainability in ML models.





Terminology clarification

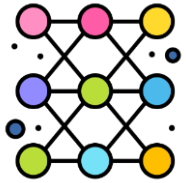
we clarify the distinction and similarities among terms often used in the ethical AI and XAI communities.

- **Understandability**(or equivalently, **intelligibility**): denotes the characteristic of a model to make a human understand its function how the model works without any need for explaining its internal structure or the algorithmic means by which the model processes data internally.

- **Comprehensibility** : When conceived for ML models, comprehensibility refers to the ability of a learning algorithm to represent its learned knowledge in a human understandable fashion.

- **Interpretability** : It is defined as the ability to explain or to provide the meaning in understandable terms to a human.

- **Explainability** : Explainability is associated with the notion of explanation as an interface between humans and a decision maker that is, at the same time, both an accurate proxy of the decision maker and comprehensible to humans.



Terminology clarification

- **Transparency** : A model is considered to be transparent if by itself it is understandable. Since a model can feature different degrees of understandability.

transparent models in Section 3 are divided into three categories:

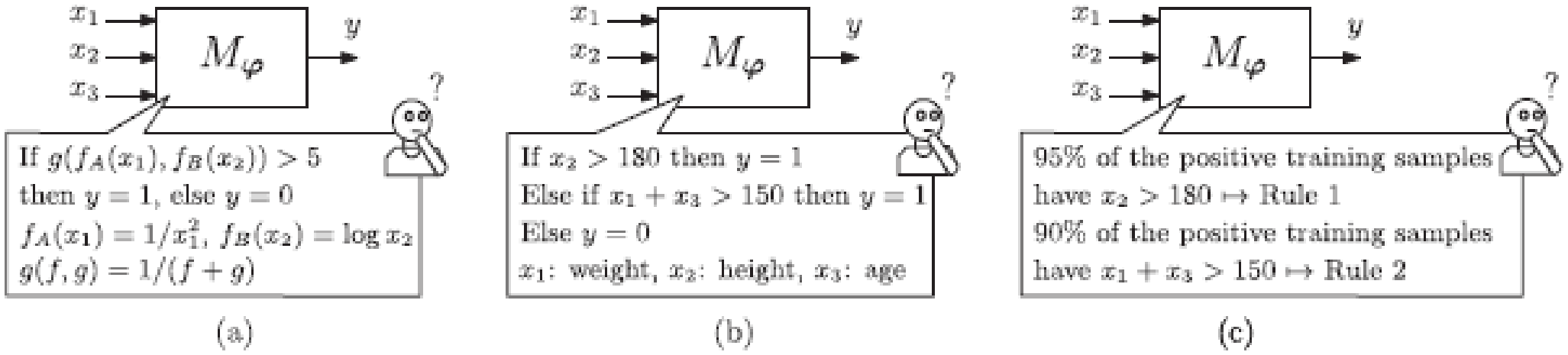
- **Simulatable models** : denotes the ability of a model of being simulated or thought about strictly by a human, hence complexity takes a dominant place in this class.

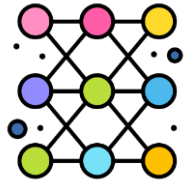
- **Decomposable models** : stands for the ability to explain each of the parts of a model (input, parameter and calculation).

- **Algorithmic transparency** : can be seen in different ways. It deals with the ability of the user to understand the process followed by the model to produce any given output from its input data.

Terminology clarification

Conceptual diagram exemplifying the different levels of transparency characterizing a ML model.





Terminology clarification

- **XAI (new definition):** *Given an audience, an **explainable** Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand.*

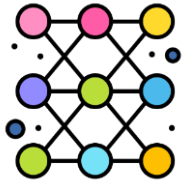
We now synthesize and enumerate definitions for these XAI goals, so as to settle a first classification criteria for the full suit of papers covered in this review:

- **Trustworthiness:** Trustworthiness might be considered as the confidence of whether a model will act as intended when facing a given problem.

- **Causality:** Another common goal for explainability is that of finding causality among data variables.

- **Transferability**: Models are always bounded by constraints that should allow for their seamless transferability.

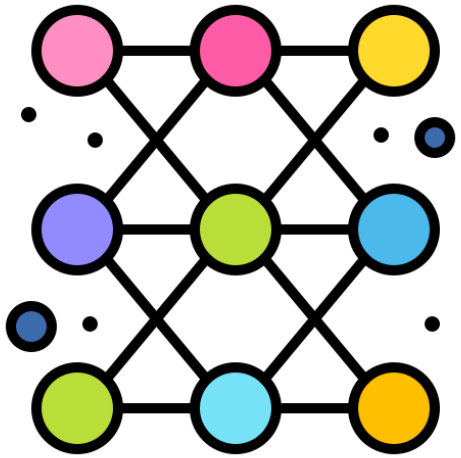
- **Informativeness:** explainable ML models should give infor- mation about the problem being tackled.



Terminology clarification

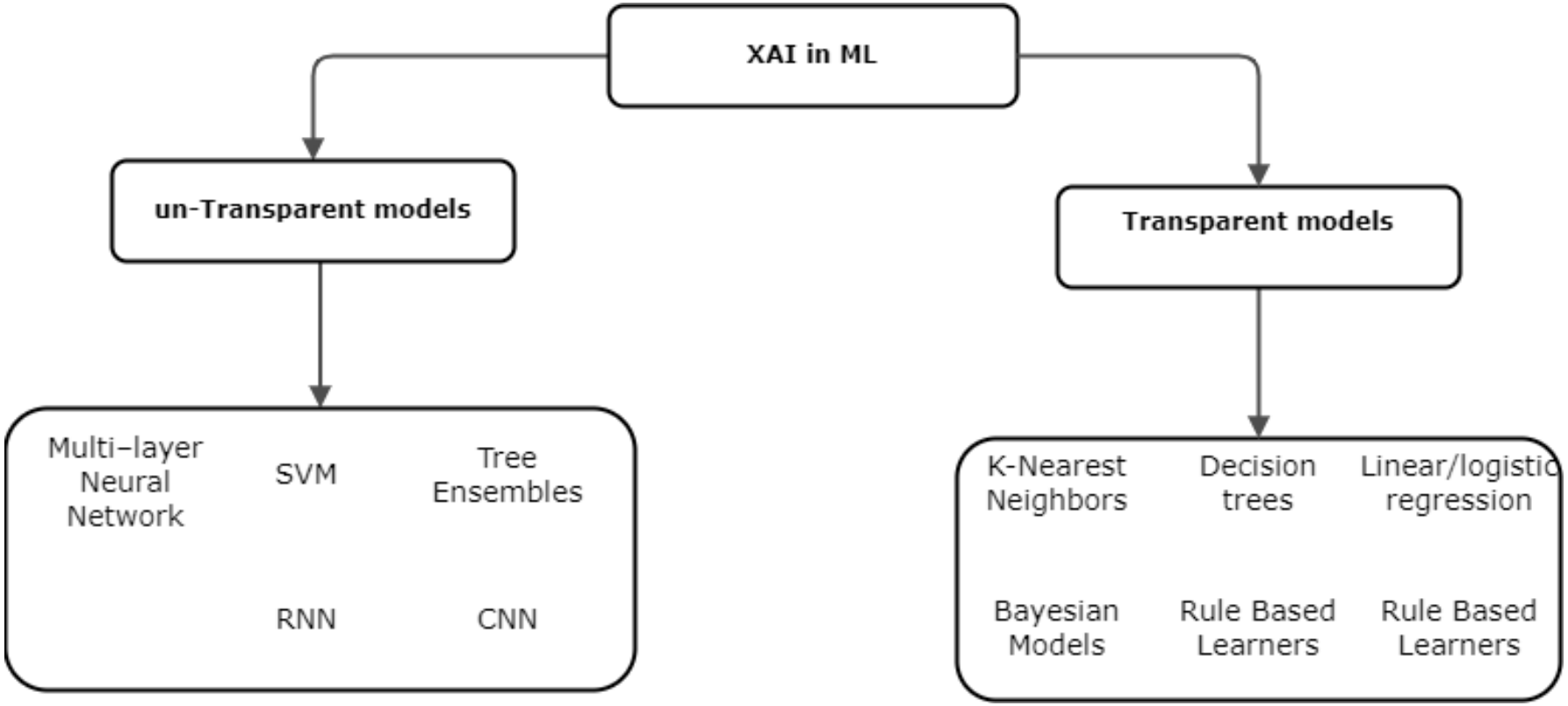
Goals pursued in the reviewed literature toward reaching explainability, and their main target audience.

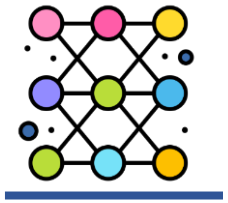
XAI Goal	Main target audience (Fig. 2)
Trustworthiness	Domain experts, users of the model affected by decisions
Causality	Domain experts, managers and executive board members, regulatory entities/agencies
Transferability	Domain experts, data scientists
Informativeness	All
Confidence	Domain experts, developers, managers, regulatory entities/agencies
Fairness	Users affected by model decisions, regulatory entities/agencies
Accessibility	Product owners, managers, users affected by model decisions
Interactivity	Domain experts, users affected by model decisions
Privacy awareness	Users affected by model decisions, regulatory entities/agencies



XAI in ML models

Taxonomy of models

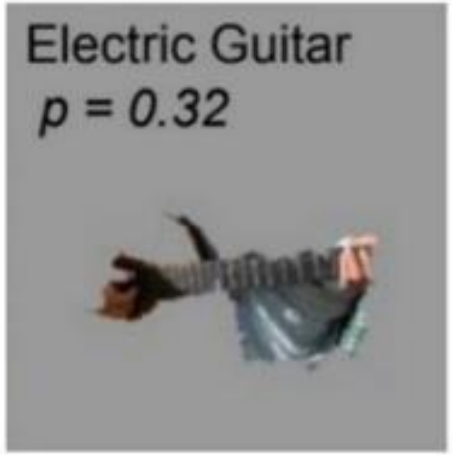




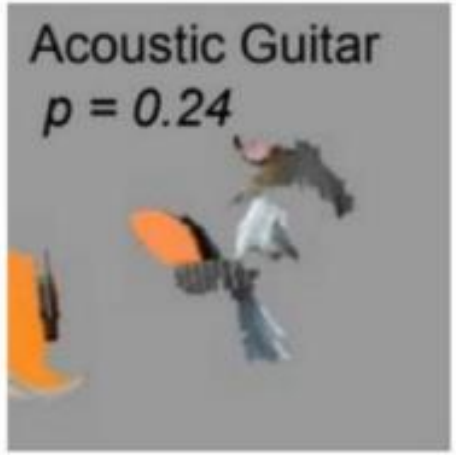
Model	Transparent ML Models			Post-hoc analysis
	Simulatability	Decomposability	Algorithmic Transparency	
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in them have grown to force decomposition	Variables and interactions are too complex to be analyzed without mathematical tools	Not needed
Decision Trees	A human can simulate and obtain the prediction of a decision tree on his/her own, without requiring any mathematical background	The model comprises rules that do not alter data whatsoever, and preserves their readability	Human-readable rules that explain the knowledge learned from data and allows for a direct understanding of the prediction process	Not needed
K-Nearest Neighbors	The complexity of the model (number of variables, their understandability and the similarity measure under use) matches human naive capabilities for simulation	The amount of variables is too high and/or the similarity measure is too complex to be able to simulate the model completely, but the similarity measure and the set of variables can be decomposed and analyzed separately	The similarity measure cannot be decomposed and/or the number of variables is so high that the user has to rely on mathematical and statistical tools to analyze the model	Not needed
Rule Based Learners	Variables included in rules are readable, and the size of the rule set is manageable by a human user without external help	The size of the rule set becomes too large to be analyzed without decomposing it into small rule chunks	Rules have become so complicated (and the rule set size has grown so much) that mathematical tools are needed for inspecting the model behaviour	Not needed
General Additive Models	Variables and the interaction among them as per the smooth functions involved in the model must be constrained within human capabilities for understanding	Interactions become too complex to be simulated, so decomposition techniques are required for analyzing the model	Due to their complexity, variables and interactions cannot be analyzed without the application of mathematical and statistical tools	Not needed
Bayesian Models	Statistical relationships modeled among variables and the variables themselves should be directly understandable by the target audience	Statistical relationships involve so many variables that they must be decomposed in marginals so as to ease their analysis	Statistical relationships cannot be interpreted even if already decomposed, and predictors are so complex that model can be only analyzed with mathematical tools	Not needed
Tree Ensembles	✗	✗	✗	Needed: Usually *Model simplification* or *Feature relevance* techniques
Support Vector Machines	✗	✗	✗	Needed: Usually *Model simplification* or *Local explanations* techniques
Multi–layer Neural Network	✗	✗	✗	Needed: Usually *Model simplification*, *Feature relevance* or *Visualization* techniques
Convolutional Neural Network	✗	✗	✗	Needed: Usually *Feature relevance* or *Visualization* techniques
Recurrent Neural Network	✗	✗	✗	Needed: Usually *Feature relevance* techniques

Post-hoc Explainability technique

(a) Original image



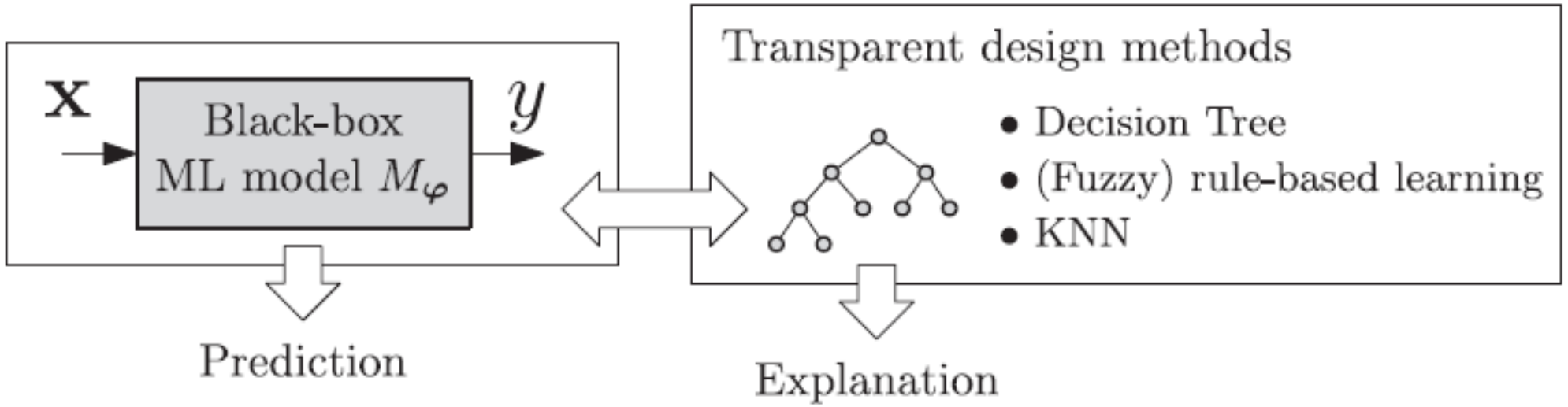
(b) Explaining *electric guitar*

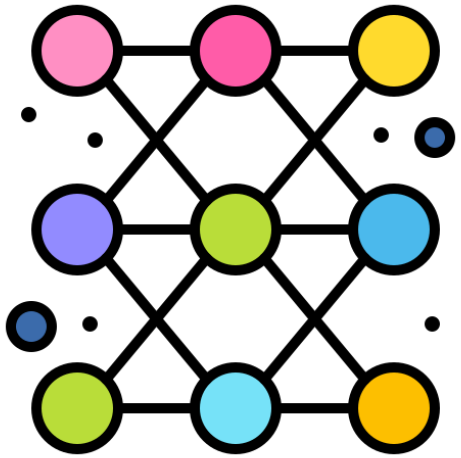


(c) Explaining *acoustic guitar*

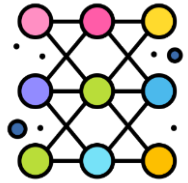
Hybrid transparent and black-box methods

Pictorial representation of a hybrid model. A neural network considered as a black-box can be explained by associating it to a more interpretable model such as a Decision Tree, a (fuzzy) rule-based system or KNN





XAI: Opportunities, challenges and future research needs



research opportunities

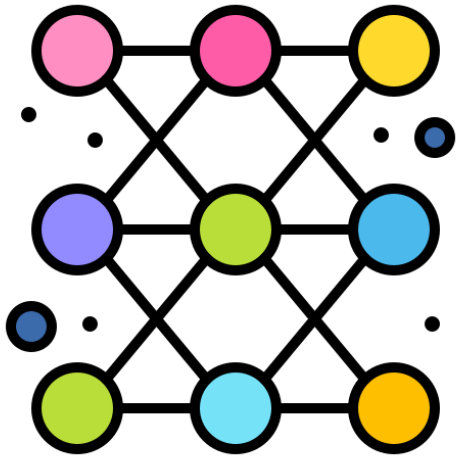
Identifying possible research paths that can be followed to address them effectively in years to come:

- we will stress on the potential of XAI developments to effectively achieve an optimal balance between the interpretability and performance of ML models.

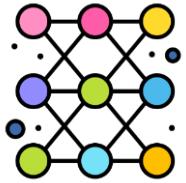
- we stressed on the imperative need for reaching a consensus on *what* explainability entails within the AI realm. Reasons for pursuing explainability are also assorted and, under our own assessment of the literature so far, not unambiguously mentioned throughout related works.

- Deep Learning models, examining advances reported so far around a specific bibliographic taxonomy.

- we close up this prospective discussion, which place on the table several research niches that despite its connection to model explainability, remain insufficiently studied by the community.



Responsible AI



Principles

the authors show a visual frame- work where different organizations are classified according to the following parameters:

- Nature, which could be private sector, government, inter- governmental organization, civil society or multistakeholder.

- Content of the principles: eight possible principles such as privacy, explainability, or fairness. They also consider the coverage that the document grants for each of the considered principles.

- Target audience: to whom the principles are aimed. They are normally for the organization that developed them, but they could also be destined for another audience.

- Whether or not they are rooted on the International Human Rights, as well as whether they explicitly talk about them.

Principles

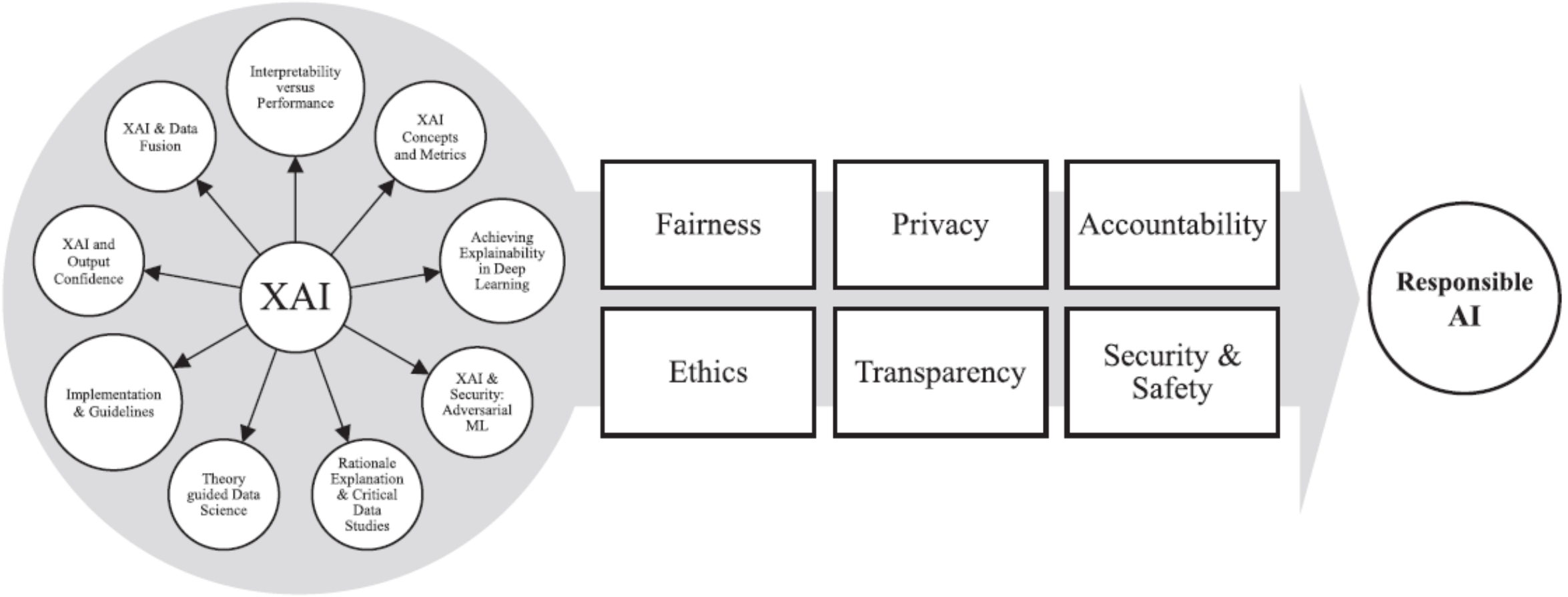
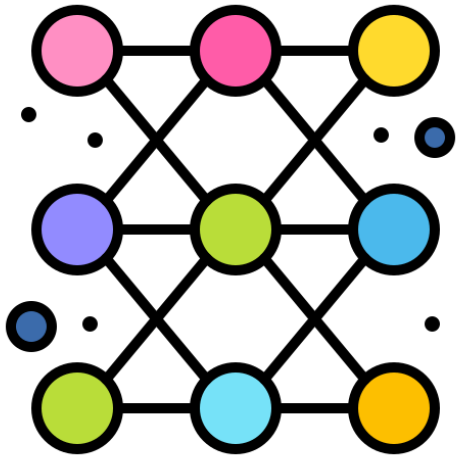
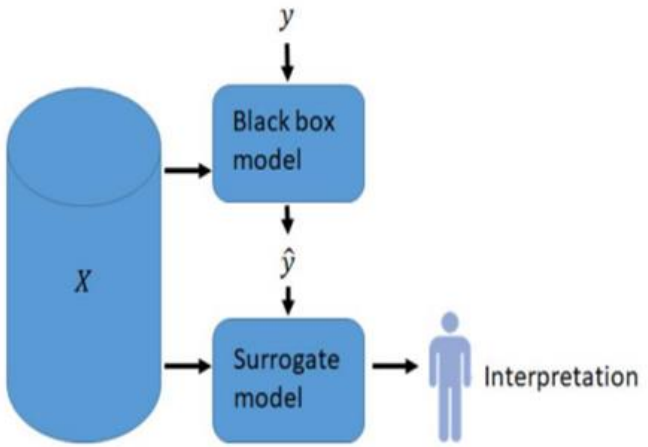
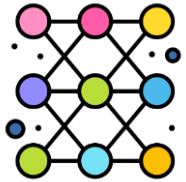


Fig. 14. Summary of XAI challenges discussed in this overview and its impact on the principles for Responsible AI.

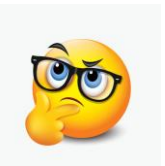


Lime = Local Interpretable model-agnostic explanations





Why should I trust you?



the recipe for training local surrogate models is as follows:

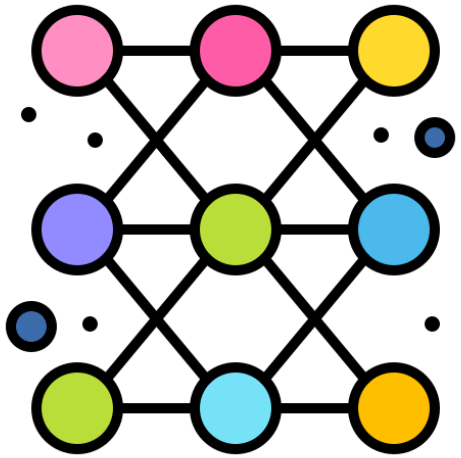
•1 Select your instance of interest for which you want to have an explanation of its black box prediction.

•2 Perturb your dataset and get the black box predictions for these new points.

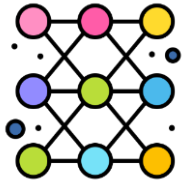
•3 Weight the new samples according to their proximity to the instance of interest.

•4 Train a weighted, interpretable model on the dataset with the variations.

•5 Explain the prediction by interpreting the local model.



Conclusions and outlook



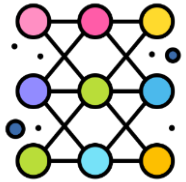
Conclusions and outlook

We have moved our discussions beyond what has been made so far in the XAI realm toward the concept of Responsible AI, a paradigm that imposes a series of AI principles to be met when implementing AI models in practice, including fairness, transparency, and privacy.

We have also discussed the implications of adopting XAI techniques in the context of data fusion, unveiling the potential of XAI to compromise the privacy of protected data involved in the fusion process.

Implications of XAI in fairness have also been discussed in detail.

This vision of XAI as a core concept to ensure the aforementioned principles for Responsible AI is summarized graphically in Fig.



references

1) Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R. and Chatila, R., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, 58, pp.82-115.

2) Ribeiro, M.T., Singh, S. and Guestrin, C., 2016, August. " Why should i trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144).