Tópicos selectos de computación

María del Rocío Ochoa Montiel

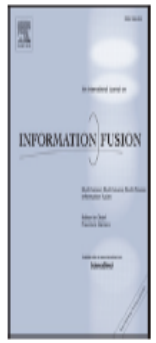
Contacto: ma.rocio.ochoa@gmail.com

Information Fusion 58 (2020) 82–115

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus



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

Alejandro Barredo Arrieta[a], Natalia Díaz-Rodríguez[b], Javier Del Ser[a,c,d,*], Adrien Bennetot[b,e,f], Siham Tabik[g], Alberto Barbado[h], Salvador Garcia[g], Sergio Gil-Lopez[a], Daniel Molina[g], Richard Benjamins[h], Raja Chatila[f], Francisco Herrera[g]

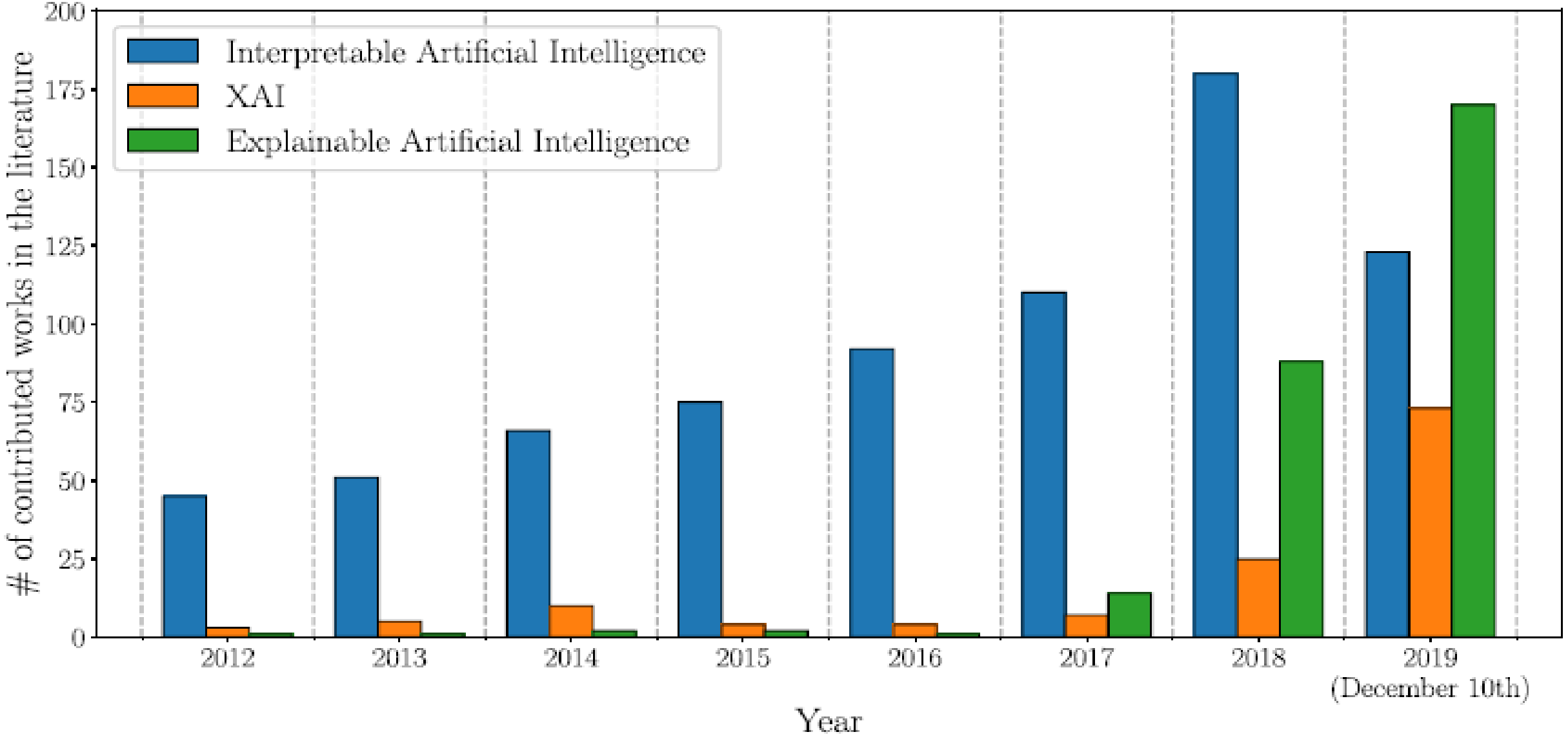
[a] TECNALIA, Derio 48160, Spain
[b] ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France
[c] University of the Basque Country (UPV/EHU), Bilbao 48013, Spain
[d] Basque Center for Applied Mathematics (BCAM), Bilbao 48009, Bizkaia, Spain
[e] Segula Technologies, Parc d'activité de Pissaloup, Trappes, France
[f] Institut des Systèmes Intelligents et de Robotique, Sorbonne Université, France
[g] DaSCI Andalusian Institute of Data Science and Computational Intelligence, University of Granada, Granada 18071, Spain
[h] Telefonica, Madrid 28050, Spain

DARPA's Explainable Artificial Intelligence Program

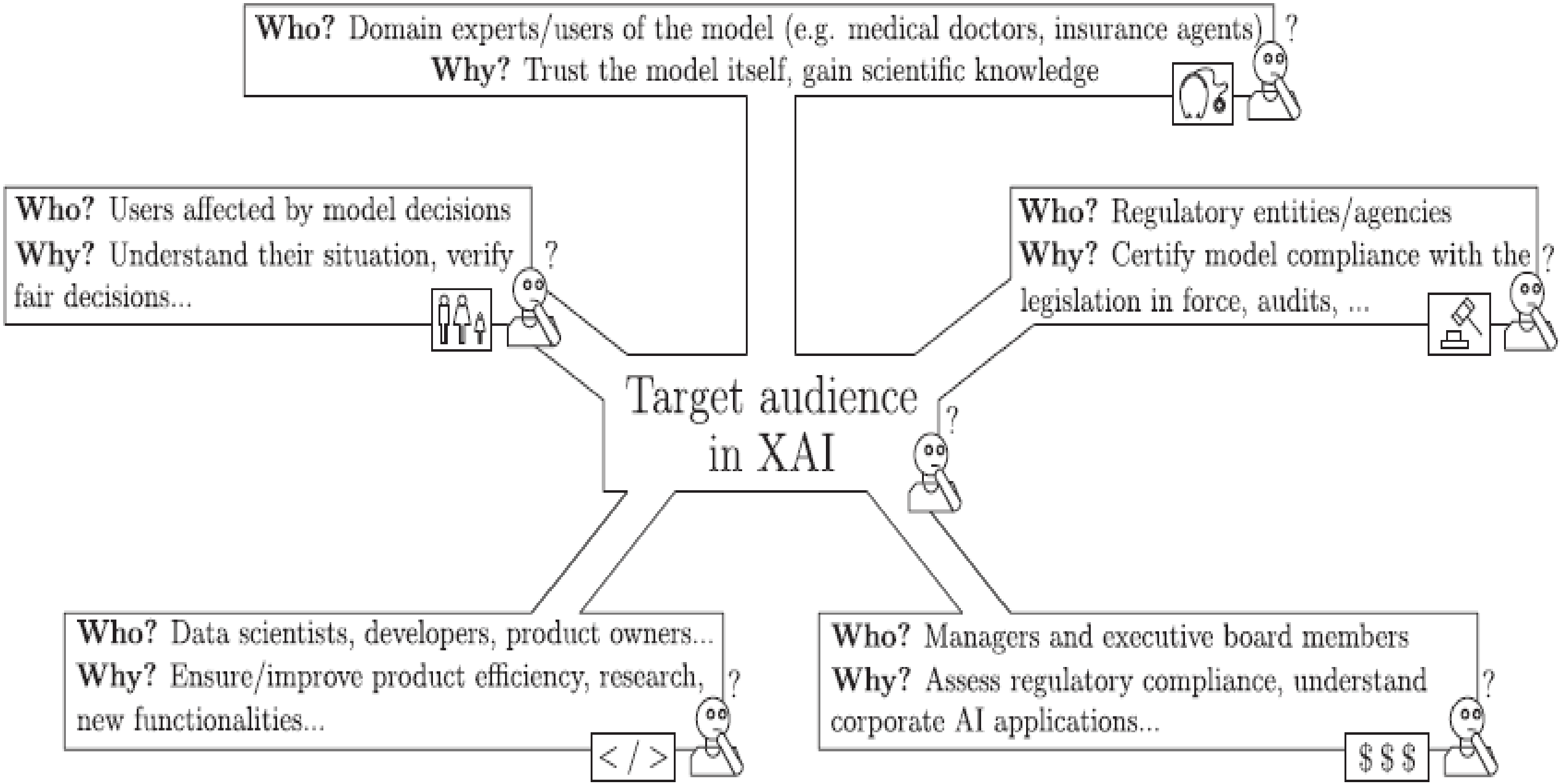
David Gunning, David W. Aha

 ISSN 0738-460

Evolution of XAI (eXplainable AI)



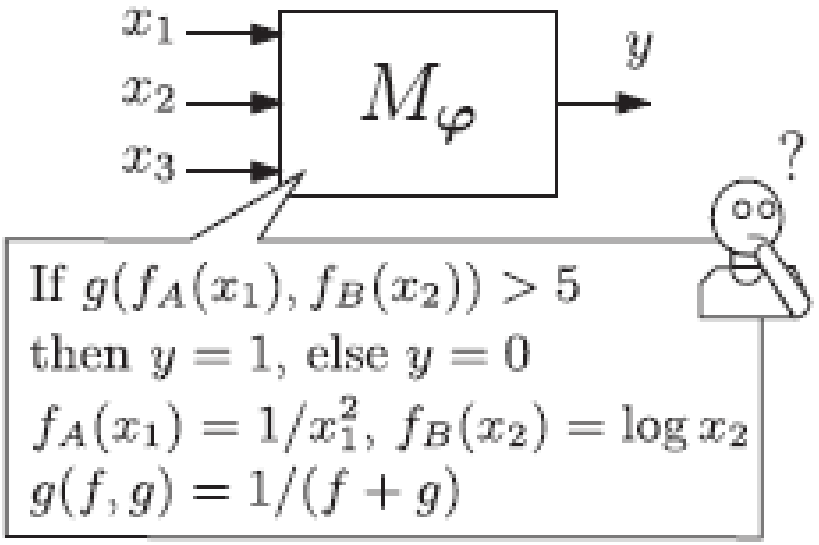
Purposes of explainability in ML(machine learning) models



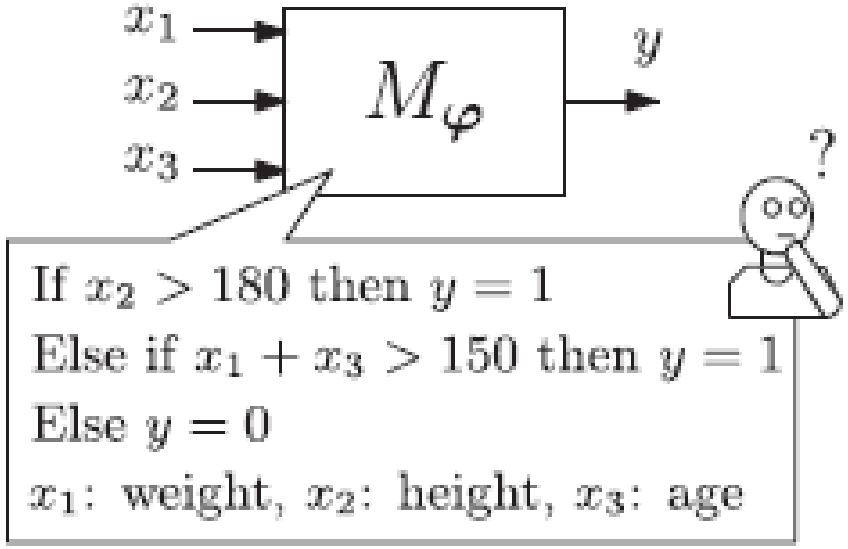
Goals pursued in the reviewed literature toward reaching explainability, and their main target audience

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

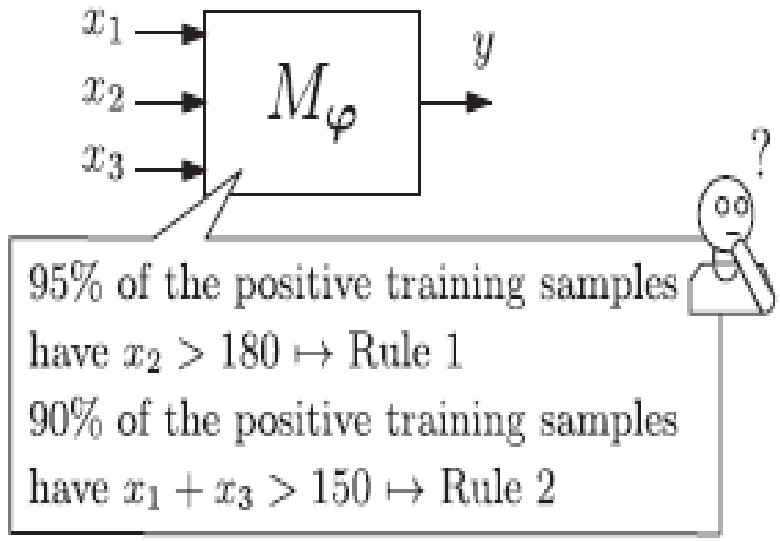
Levels of transparency characterizing a ML model



a) simulatability



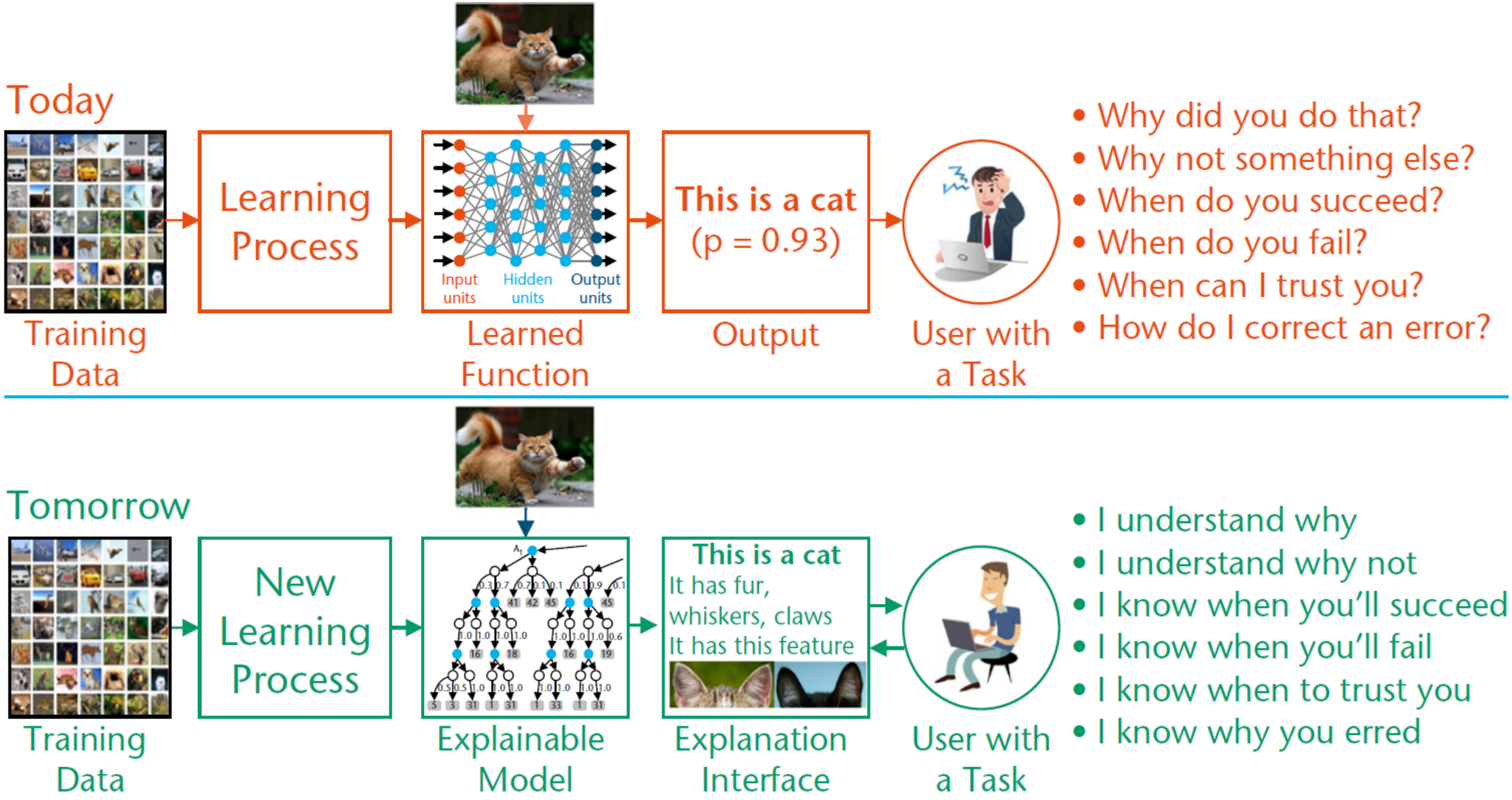
b) decomposability



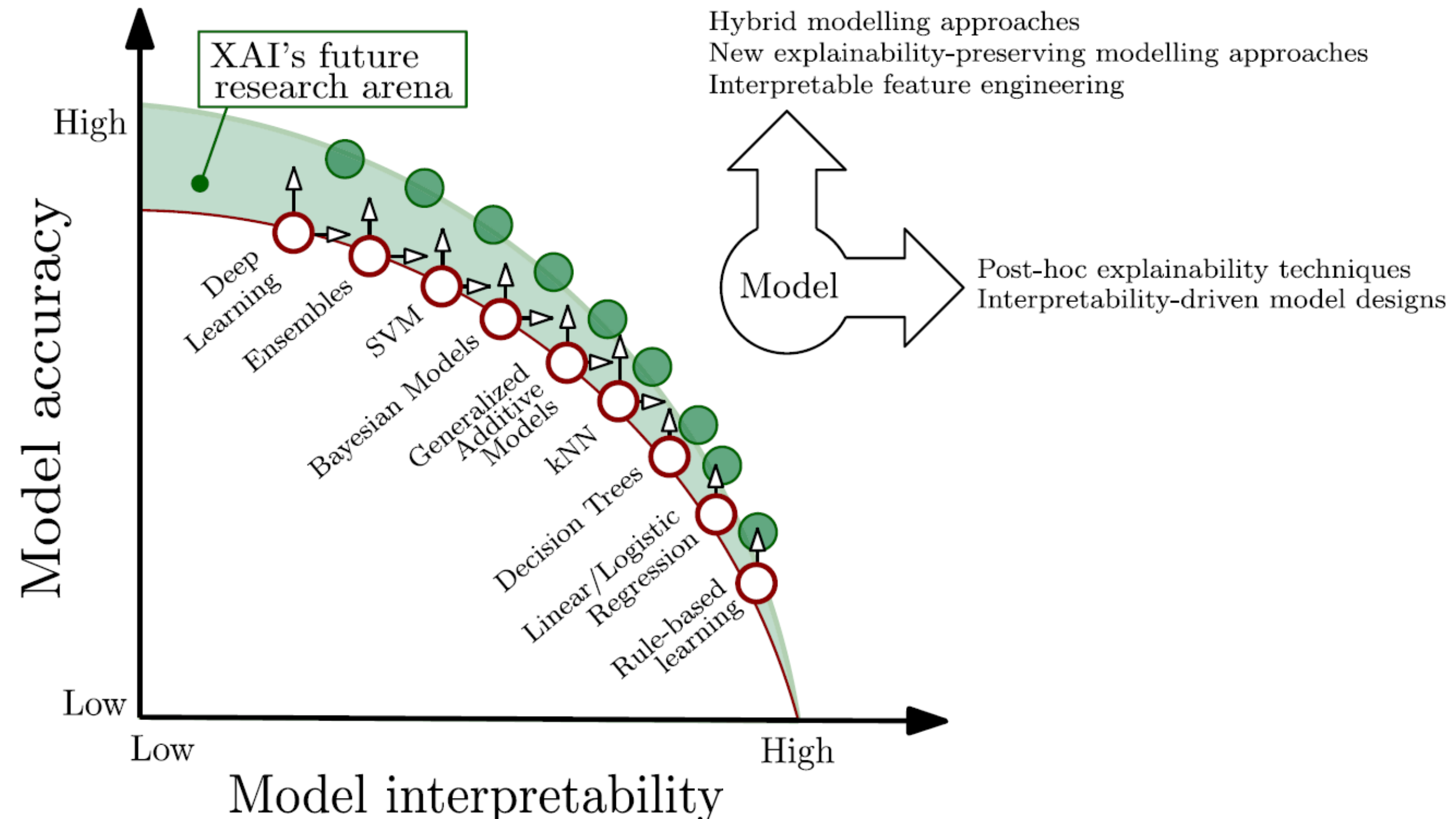
c) Algorithmic transparency

Targets for explainability may include:

- a given example,
- the output classes, or
- the dataset itself



Gunning, D., & Aha, D. (2019). DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), 44-58. https://doi.org/10.1609/aimag.v40i2.2850.


XAI's future research arena
Hybrid modelling approaches
New explainability-preserving modelling approaches
Interpretable feature engineering
Model
Post-hoc explainability techniques
Interpretability-driven model designs
Deep Learning
Ensembles
SVM
Bayesian Models
Generalized Additive Models
kNN
Decision Trees
Linear/Logistic Regression
Rule-based learning
High
Low
Low
High
Model accuracy
Model interpretability

Post-hoc explainability approaches for a ML model