Tema 12

Ética e explicabilidade

Professora:
Ariane Machado Lima

SISTEMAS DE INFORMAÇÃO

EACH

O problema...

Qual é o problema nos modelos de aprendizagem de máquina?

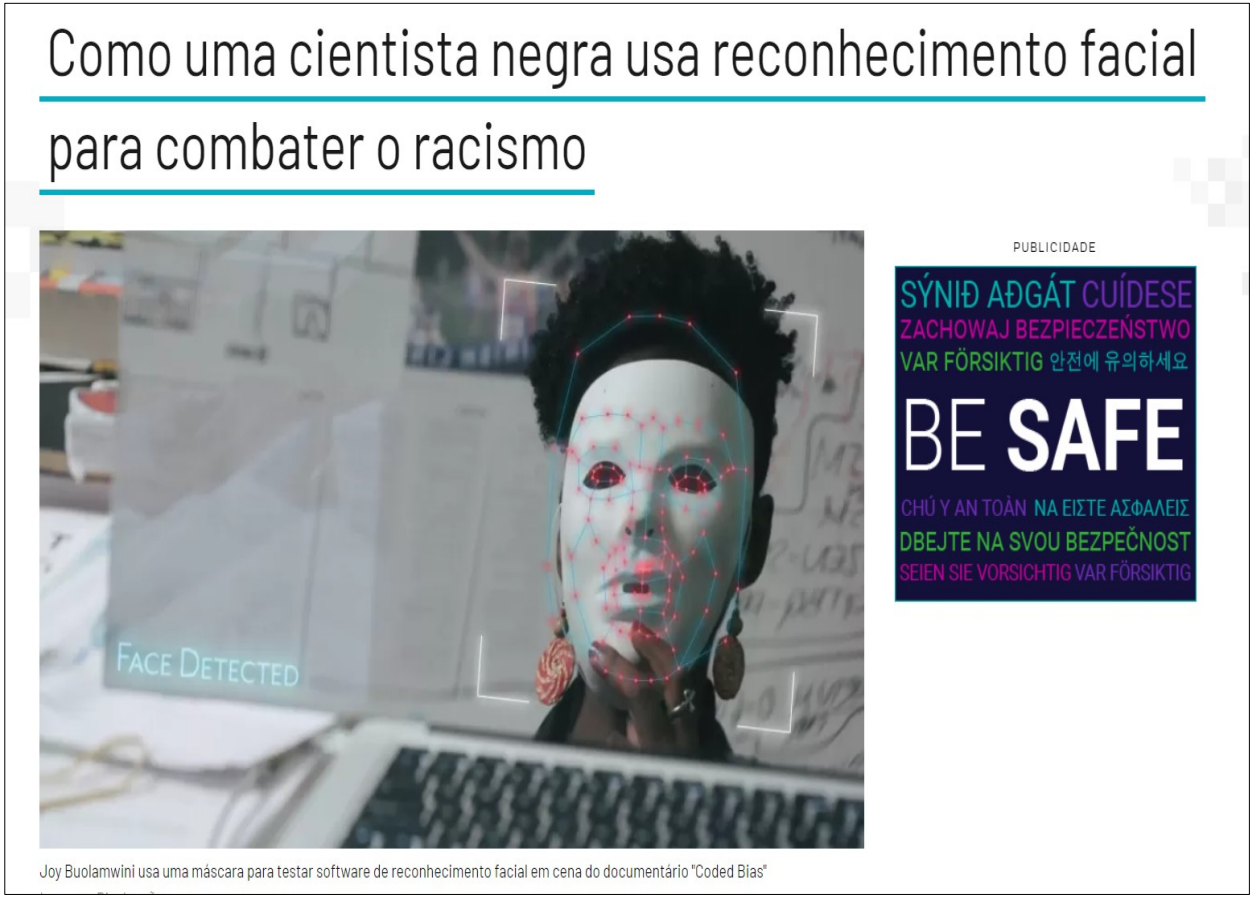
- Não é a popularidade dos modelos de IA que é a problemática, é se eles são justos e podem ser **<u>responsabilizados</u>**

- Foram documentadas uma longa lista de IAs que tomaram decisões ruins por causa de: erros de codificação, preconceitos arraigados nos dados em que foram treinados os modelos, etc.

- Exemplos:

https://brasil.elpais.com/brasil/2018/01/14/tecnologia/1515955554_803955.html

- Exemplos:



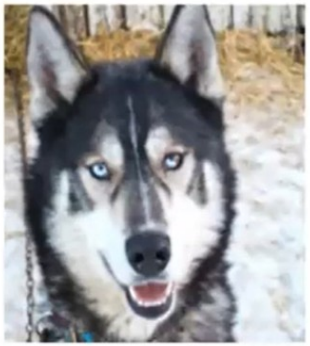
https://www.uol.com.br/tilt/colunas/akin-abaz/2021/02/11/joy-buolamwini-a-luta-contra-o-racismo-atraves-dos-algoritmos.htm

O problema de modelos de IA

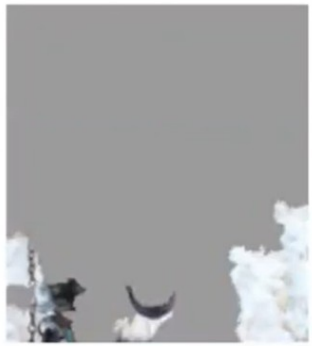
- Desconfiança:
  - será que está certo mesmo?
  - teria vieses? (étnicos, gênero, nacionalidade, etc)
  - acertando pelo motivo errado? (ex: lobos)
  - “Por que eu fui prejudicado? Mereço saber!”
- Alguns modelos respondem parte dessas perguntas, mas muitos (os mais novos – deep learning) nem isso
- Necessidade de IA confiável

Será que a explicação de um modelo muda a percepção e confiança das pessoas ?

Sabendo como esse modelo funciona muda como as pessoas se relacionam com o modelo.



(a) Husky classified as wolf



(b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3391288

Publicações

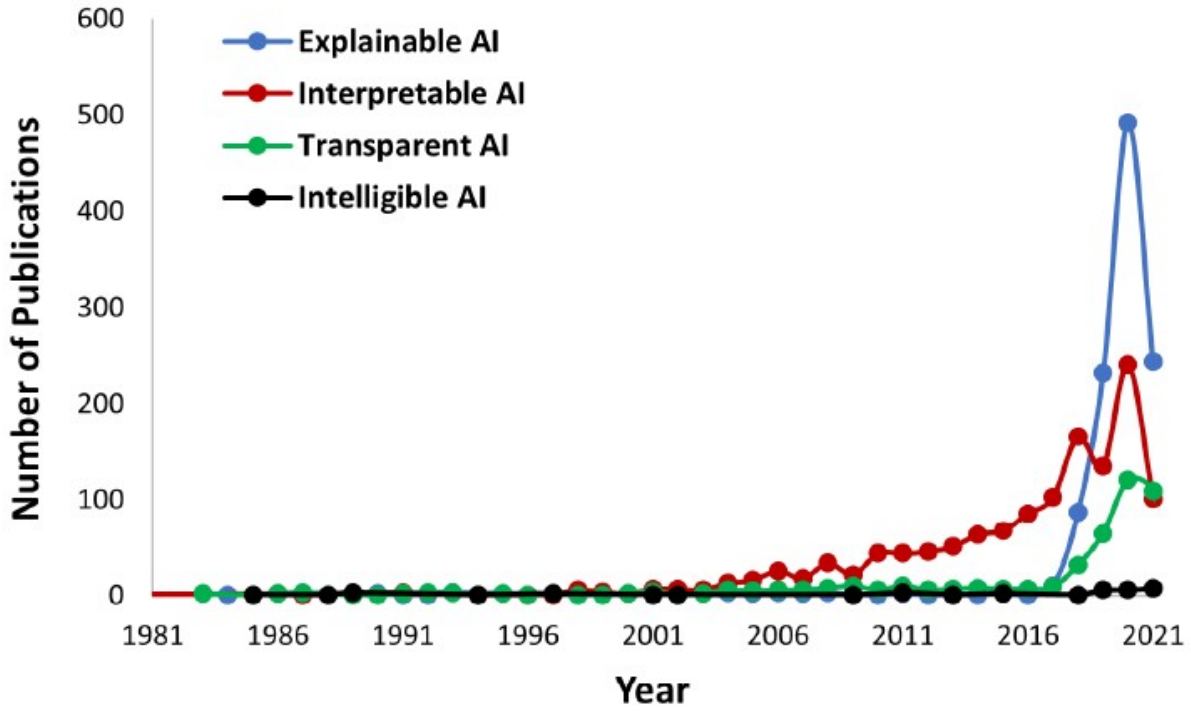


Fig. 1: Yearly publications for Explainable, Interpretable, Transparent and Intelligible AI. (Data derived from SCOPUS)

IA Responsável

2021 IEEE 25th International Enterprise Distributed Object Computing Workshop (EDOCW)

A Survey on Explainable Artificial Intelligence Techniques and Challenges

Ambreen Hanif
Department of Computing
Macquarie University
Sydney, Australia
ambreen.hanif@hdr.mq.edu.au

Xuyun Zhang
Department of Computing
Macquarie University
Sydney, Australia
xuyun.zheng@mq.edu.au

Steven Wood
Prospa
Sydney, Australia
steven.wood@prospa.com

***Abstract*—In the last decade, the world has envisioned tremendous growth in technology with improved accessibility of data, cloud-computing resources, and the evolution of machine learning (ML) algorithms. The intelligent system has achieved significant performance with this growth. The state-of-the-art performance of these algorithms in various domains has increased the popularity of artificial intelligence (AI). However, alongside these achievements, the non-transparency, inscrutability and inability to expound and interpret the majority of the state-of-the-art techniques are considered an ethical issue. These flaws in AI algorithms impede the acceptance of complex ML models in a variety of fields such as medical, banking and finance, security, and education. These shortcomings have prompted many concerns about the security and safety of ML system users. These systems must be transparent, according to the current regulations and policies, in order to meet the right to explanation. Due to a lack of trust in existing ML-based systems, explainable artificial intelligence (XAI)-based methods are gaining popularity,**

in vital and sensitive areas such as healthcare, manufacturing, security, law enforcement, banking and finance, education, and construction.

Recent Machine learning (ML) algorithms have piqued the interest of investors and researchers due to occasions in which AI-based systems surpassed human specialists in a variety of open tasks. With this aspect, a few crucial systems to note are, for example, Computer vision (CV) based techniques have surpassed the conventional methods in the domain. They have conquered the human expert in related open challenges, e.g., image-classification as ImageNet [42], object-detection challenge, e.g. COCO [31] and AlexNet [27]. These accomplishments are not restricted to CV; ML approaches have also made substantial progress in natural language tasks that includes visual question answering [3] and machine translation

Nesse artigo é apresentada uma revisão sob as técnicas de XAI e são sugeridas futuras rotas de pesquisa na área de sistemas de IA responsável.

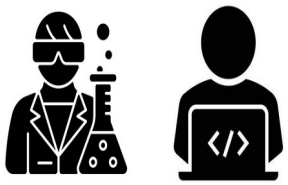
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9626294

IA Responsável

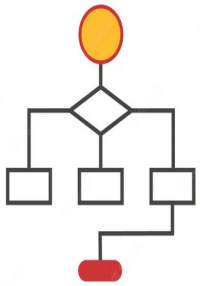
- É um conceito que representa uma mudança de paradigma de como enxergar a IA.
- Em vez de apenas pensar no **desempenho** de um modelo, os **profissionais** devem pensar também no compromisso **ético e justo** nas implicações desses modelos num contexto real.
- Entender a IA e seus **impactos** para as pessoas, impactos reais e poderosos que podem gerar benefícios ou malefícios.
- Em outras palavras, a IA Responsável é uma forma de **ensinar** as máquinas a tomar decisões de forma mais **consciente**.

IA Responsável

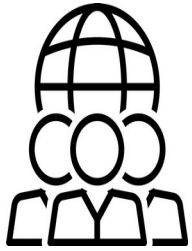
A IA responsável repousa em três pilares de igual importância (Dignum, 2017):



A sociedade deve estar preparada para **assumir a responsabilidade** pelo impacto da IA.



A necessidade de mecanismos que permitam aos próprios sistemas de IA raciocinar e agir de acordo com á ética e valores humanos.



Participação.

IA Responsável

Nesse contexto há uma discussão: caso a decisão tomada pelos sistema de IA venha a prejudicar os seres humanos, como lidar? Quem vai ser **responsável** por isso?

IA Responsável

Exemplo:

Autonomous car hits autonomous robot in bizarre collision
https://www.youtube.com/watch?v=0s4nxcleVd0

Quem será culpado se um carro autônomo machucar um pedestre?

- O construtor de hardware (dos sensores usados pelo carro)?
- O construtor de software (que permite ao carro decidir um caminho)?
- As autoridades que permitiram o uso do carro?
- O proprietário que personalizou as configurações de tomada de decisão para atender suas preferências ou não prestou atenção?
- O próprio carro porque seu comportamento é baseado em seus próprios aprendizados?

IA Responsável

Uma IA responsável leva em consideração as realidades locais; as pessoas que constroem essas IAs precisam ter pensamento crítico para avaliar os impactos a partir de princípios éticos e direitos humanos *(Global Summit on Responsible AI – Rio de Janeiro 2022)*.

IA Responsável

ITU Journal: ICT Discoveries, Special Issue No. 1, 25 Sept. 2017

RESPONSIBLE ARTIFICIAL INTELLIGENCE: DESIGNING AI FOR HUMAN VALUES

Virginia Dignum
Delft University of Technology, The Netherlands

Abstract – Artificial intelligence (AI) is increasingly affecting our lives in smaller or greater ways. In order to ensure that systems will uphold human values, design methods are needed that incorporate ethical principles and address societal concerns. In this paper, we explore the impact of AI in the case of the expected effects on the European labor market, and propose the accountability, responsibility and transparency (ART) design principles for the development of AI systems that are sensitive to human values.

Keywords – Artificial intelligence, design for values, ethics, societal impact

1. INTRODUCTION

Artificial intelligence (AI) is becoming rapidly present in all aspects of everyday life. It is everywhere, it affects everyone, and its capabilities are evolving extremely rapidly. AI can help us in systems is of the utmost relevance to AI applications such as self-driving vehicles, companion, healthcare robots, and ranking and profiling algorithms, which are already affecting society or will be in a few years. In all these applications, AI reasoning should be able

Nesse artigo são propostos “princípios de design” para o desenvolvimento de sistemas de IA: Accountability, responsabilidade e transparência (ART)

https://www.itu.int/dms_pub/itu-s/opb/journal/S-JOURNAL-ICTF.VOL1-2018-1-P01-PDF-E.pdf

Princípios ART

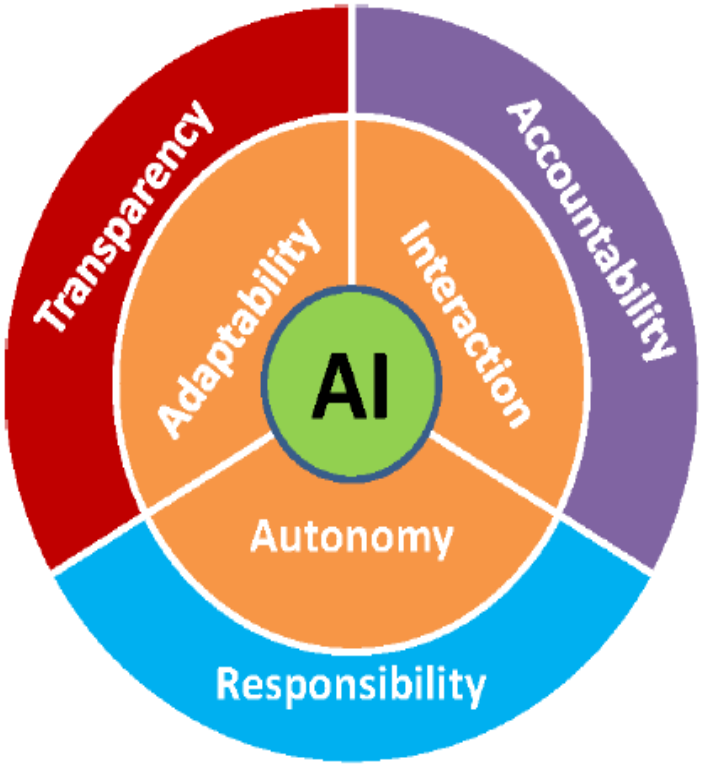


Fig. 2: The ART principles: accountability, responsibility, transparency

Para garantir que os sistemas de IA sejam desenvolvidos de forma **responsável** os autores do artigo propõem **COMPLEMENTAR** as propriedades internas pelas externas

Princípios ART

Accountability

- Necessidade de **explicar** e **justificar** as decisões e ações para quem interage com o sistema.
- Requer explicações num **contexto mais amplo** e classificando-as de acordo com os valores morais.

Responsabilidade

- É a capacidade dos sistemas de IA de **responder** pelas suas decisões e identificar erros ou resultados inesperados.

Transparency

- Refere-se à necessidade de **descrever**, **inspecionar** e **reproduzir** os **mecanismos** pelos quais os sistemas de IA tomam decisões e aprendem a se adaptar ao seu ambiente e à governança dos dados usados.

IA Explicável:

O que é explicabilidade?

Interpretabilidade ou explicabilidade?

- **Interpretabilidade**: capacidade de entender ou interpretar o funcionamento interno de um modelo, ie, como os resultados são produzidos. Modelos interpretáveis permitem que os humanos compreendam as relações entre as entradas e as saídas, facilitando a confiança e a explicação dos resultados.

- Ex: Um modelo é treinado para diagnosticar doenças com base em dados médicos. Um modelo interpretable permitiria que médicos e pesquisadores entendessem como o algoritmo chega aos diagnósticos. Por ex, se o modelo é uma árvore de decisão, é possível visualizar quais variáveis são mais importantes para a tomada de decisão.

Interpretabilidade ou explicabilidade?

- **Explicabilidade**: capacidade de explicar ou descrever o raciocínio e o processo de tomada de decisão do modelo de maneira compreensível para os seres humanos. A explicabilidade vai além da interpretabilidade, buscando fornecer razões ou justificativas para as decisões do modelo em linguagem compreensível para os humanos, muitas vezes fornecendo insights sobre por que uma determinada decisão foi tomada.
-
- Ex: o modelo interpretable diagnosticou um paciente com uma determinada condição. A explicabilidade forneceria uma explicação compreensível para o médico ou paciente sobre por que o modelo tomou essa decisão específica. Em vez de apenas mostrar as variáveis importantes, a explicabilidade poderia oferecer um resumo ou uma justificativa do diagnóstico com base nos dados do paciente, como indicadores de exames específicos, histórico médico e sintomas.

Interpretabilidade ou explicabilidade?

- No exemplo:
  - a interpretabilidade permitiria aos especialistas entenderem como o modelo chegou a uma conclusão
  - a explicabilidade proporcionaria uma narrativa compreensível sobre o raciocínio do modelo para a tomada de decisão, ajudando a construir confiança e aceitação por parte dos médicos e pacientes.

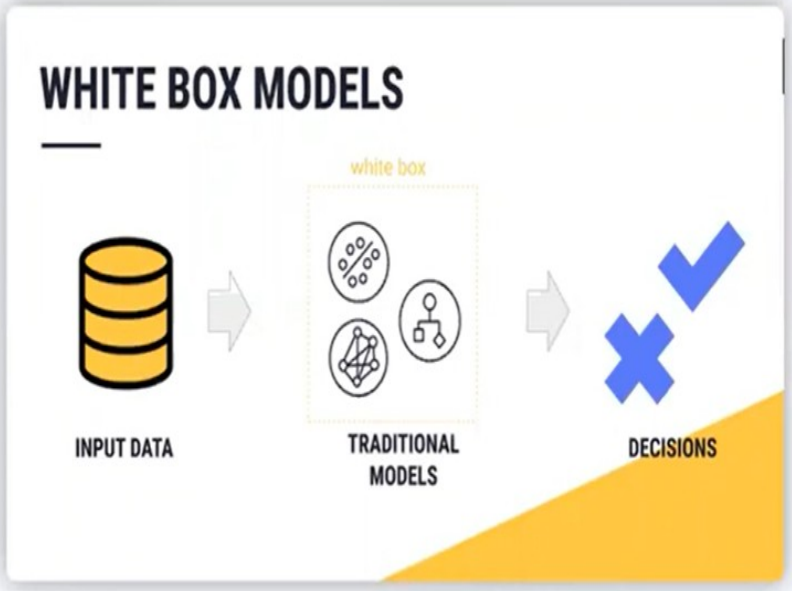
Transparência

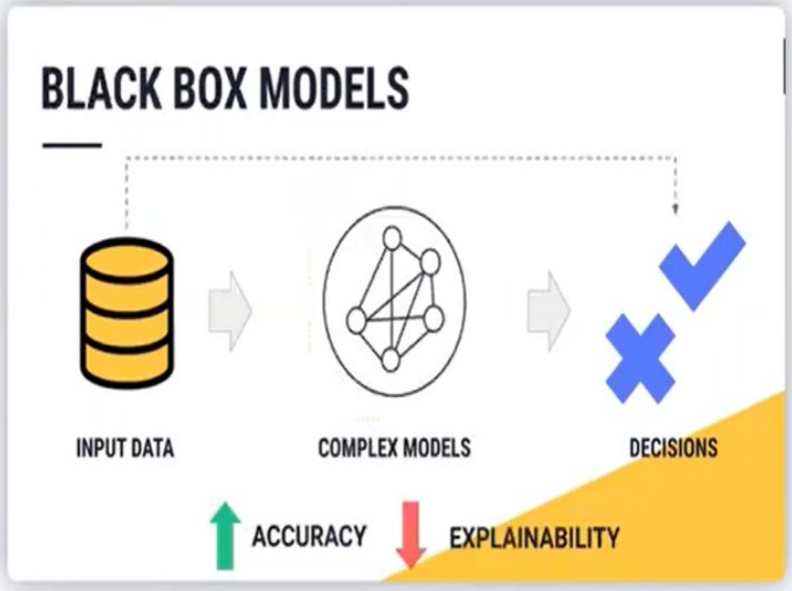
- Um modelo é transparente (o contrário de caixa-preta) quando ele por si só é interpretável (sem ter que ficar fazendo cálculos e análises sobre ele)
- Ex: árvores de decisão são transparentes, mas *random forests* nem tanto (precisa calcular as *feature importances*)

Explicabilidade é mais

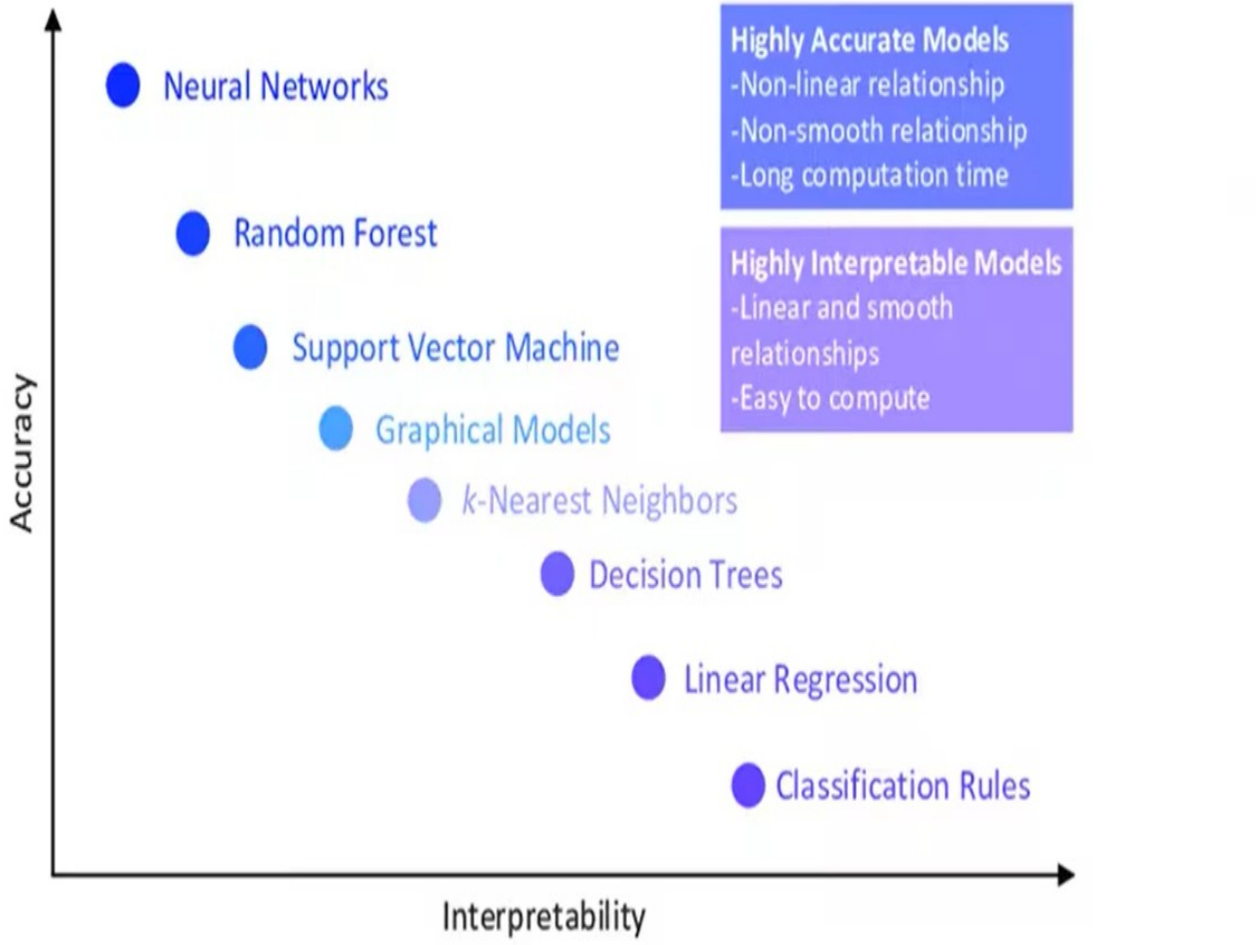
- Explicabilidade é uma atividade ativa (não passiva): você pode ter um modelo caixa-preta e ativamente aplicar um método de explicabilidade para explicar seu resultado

Acurácia x explicabilidade





Acurácia x explicabilidade

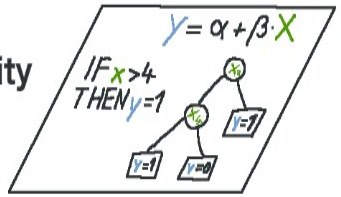


Source: Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions

Humans

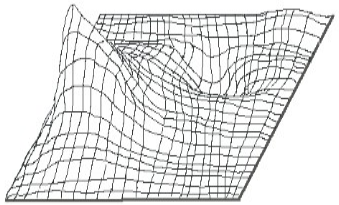
inform

Interpretability Methods



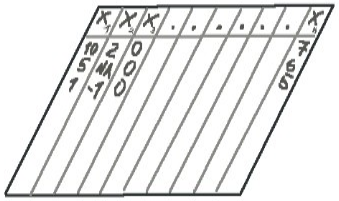
extract

Black Box Model



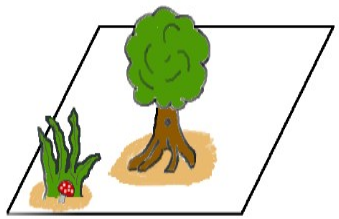
learn

Data

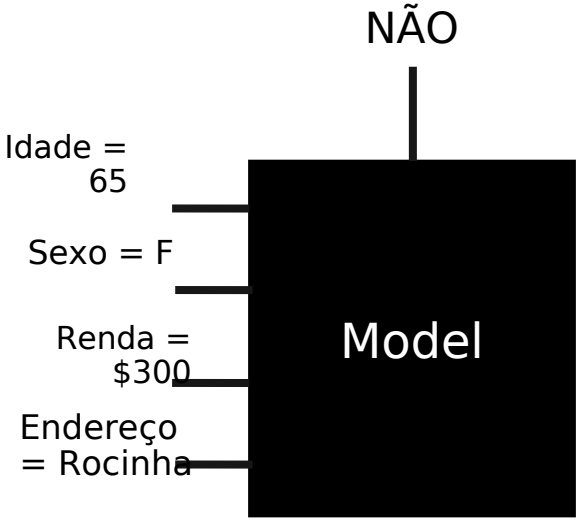


capture

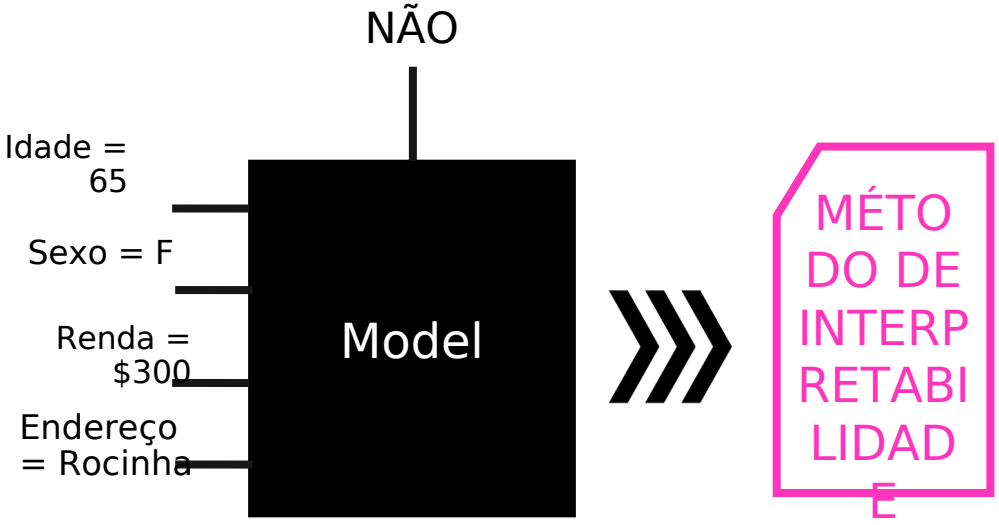
World



- O cliente teve seu pedido de crédito negado e quer saber porque?



- O cliente teve seu pedido de crédito negado e quer saber porque?



- O cliente teve seu pedido de crédito negado e quer saber porque?

- O cliente teve seu pedido de crédito negado e quer saber porque?

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion 58 (2020): 82-115
https://www.sciencedirect.com/science/article/pii/S1566253519308103?via%3Dihub

Information Fusion 58 (2020) 82–115

Contents lists available at ScienceDirect

Information Fusion

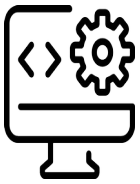
journal homepage: www.elsevier.com/locate/inffus

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

- Explicabilidade precisa levar em considerar a audiência



A compreensão do modelo

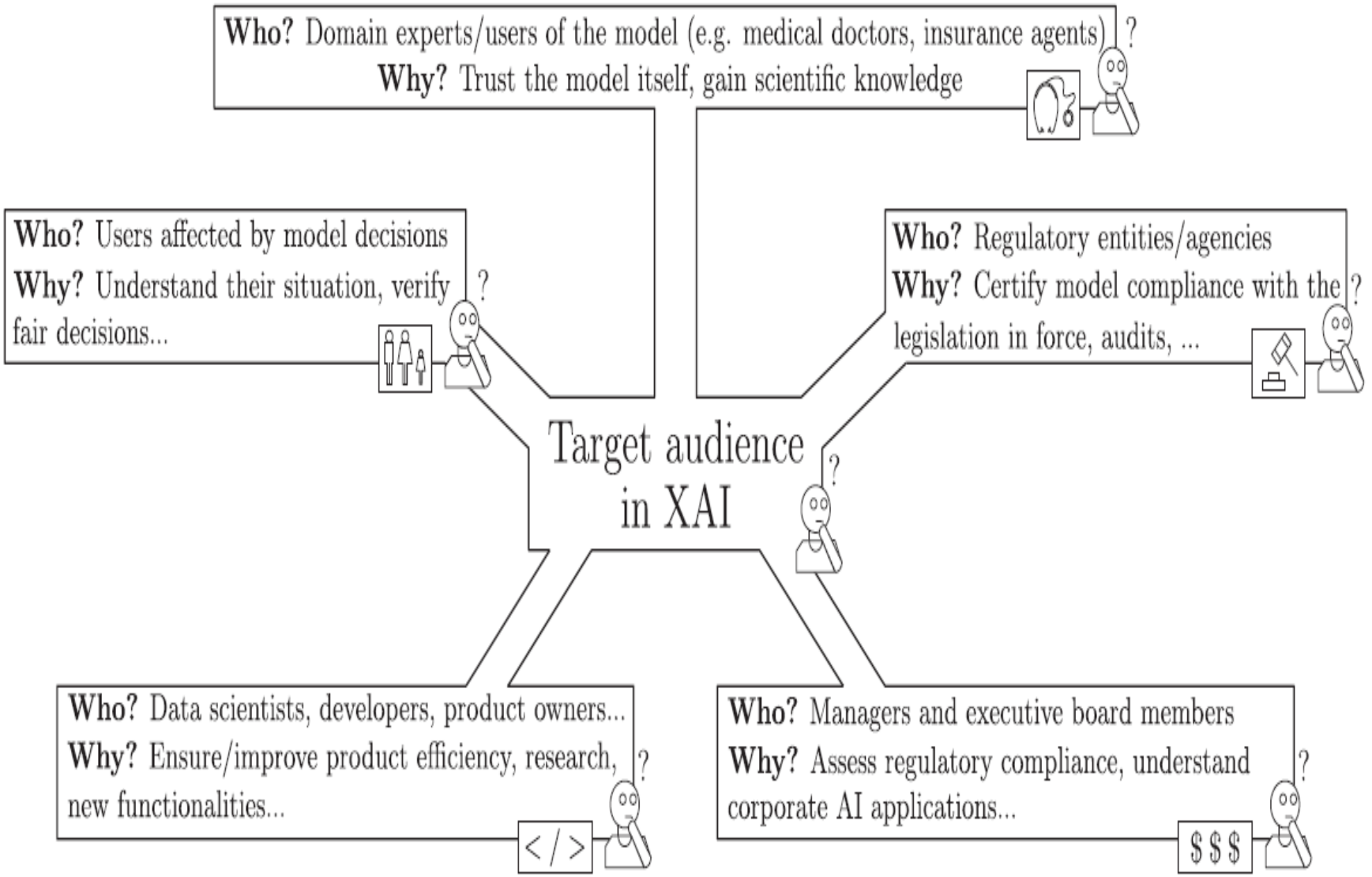
A compreensão humana

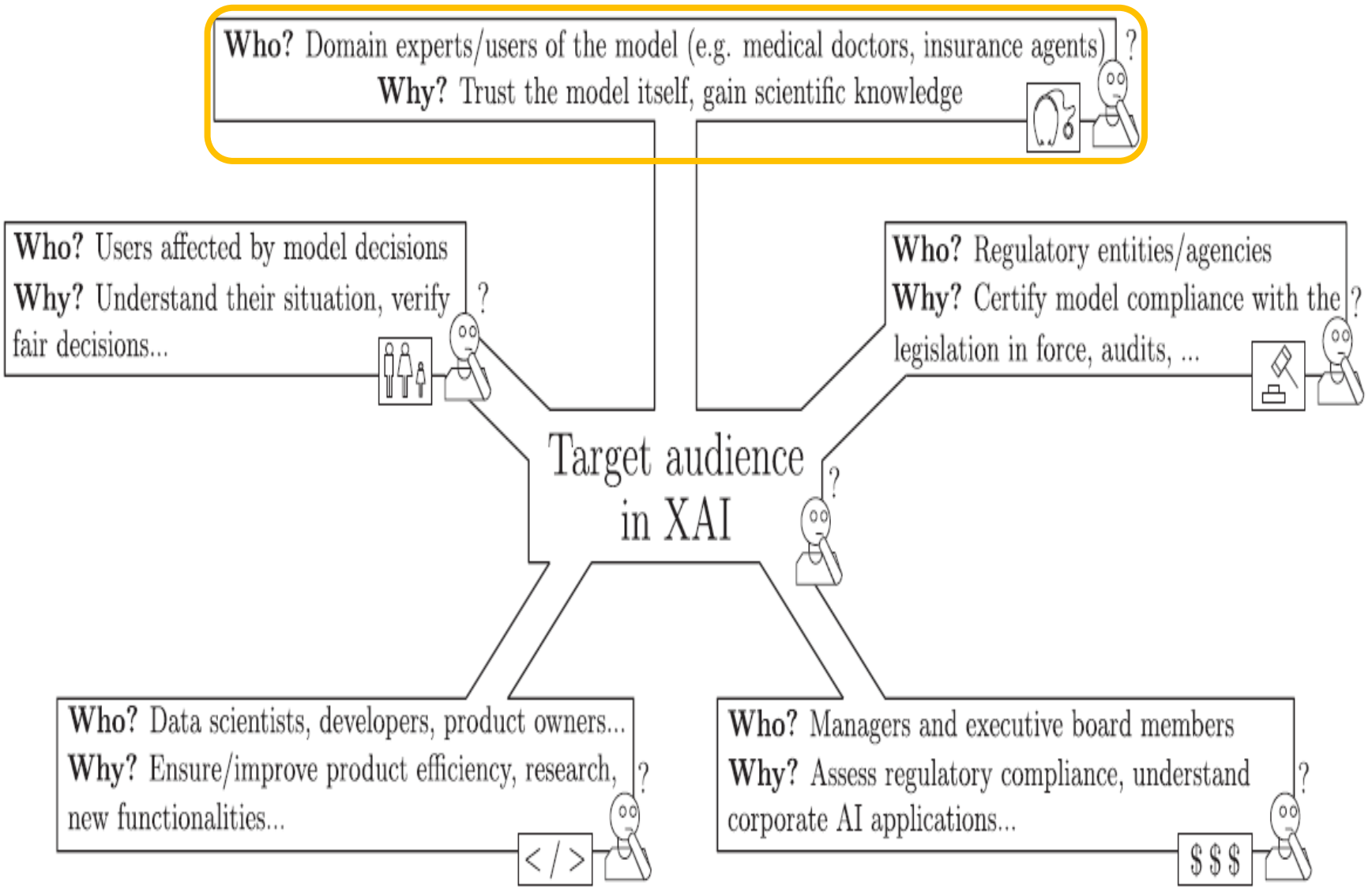
- As habilidades cognitivas e o **objetivo** dos usuários do modelo devem ser levados em consideração.

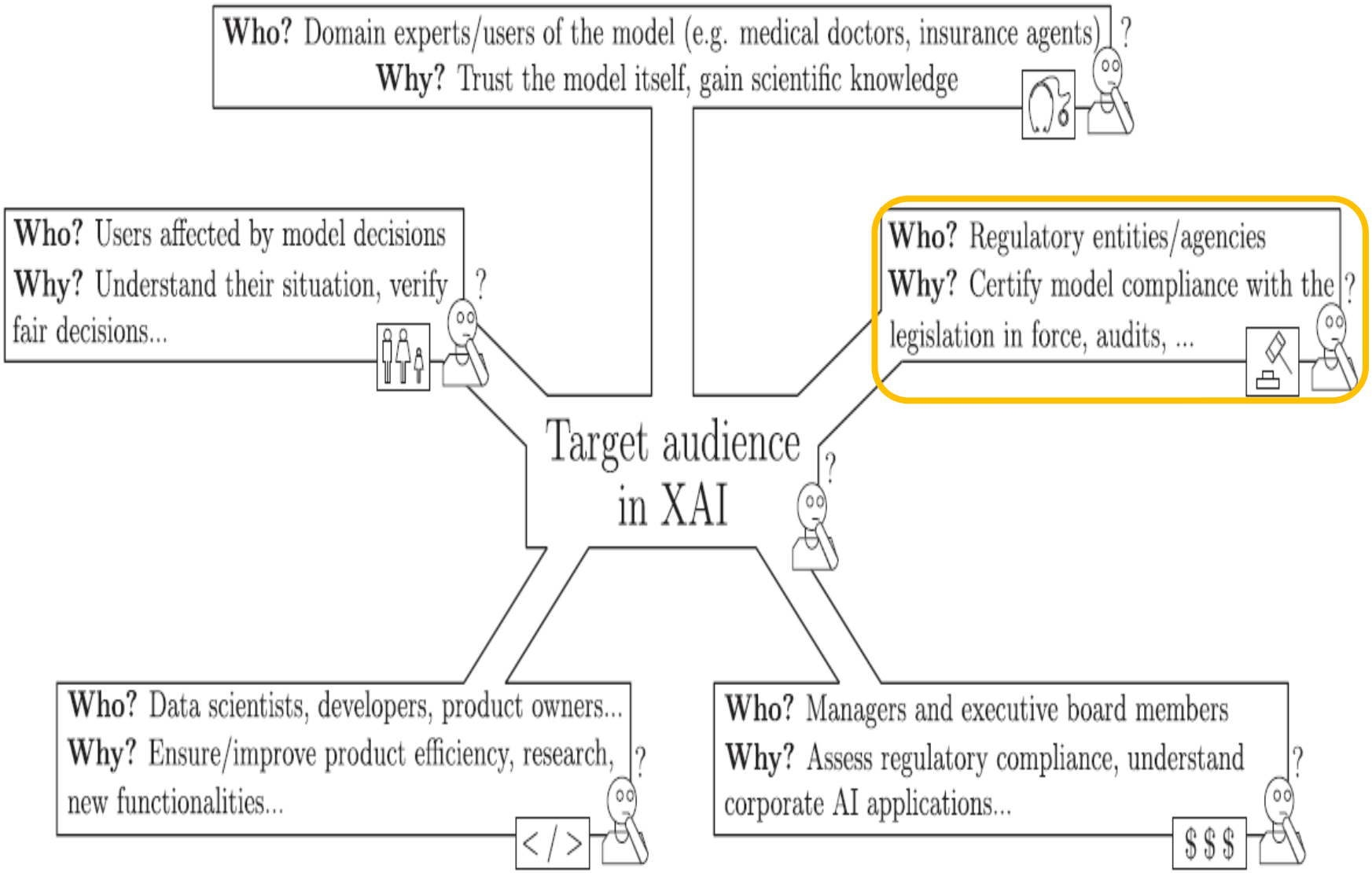
- Está é razão pela qual o conceito de **audiência** é uma pedra angular da IAX.

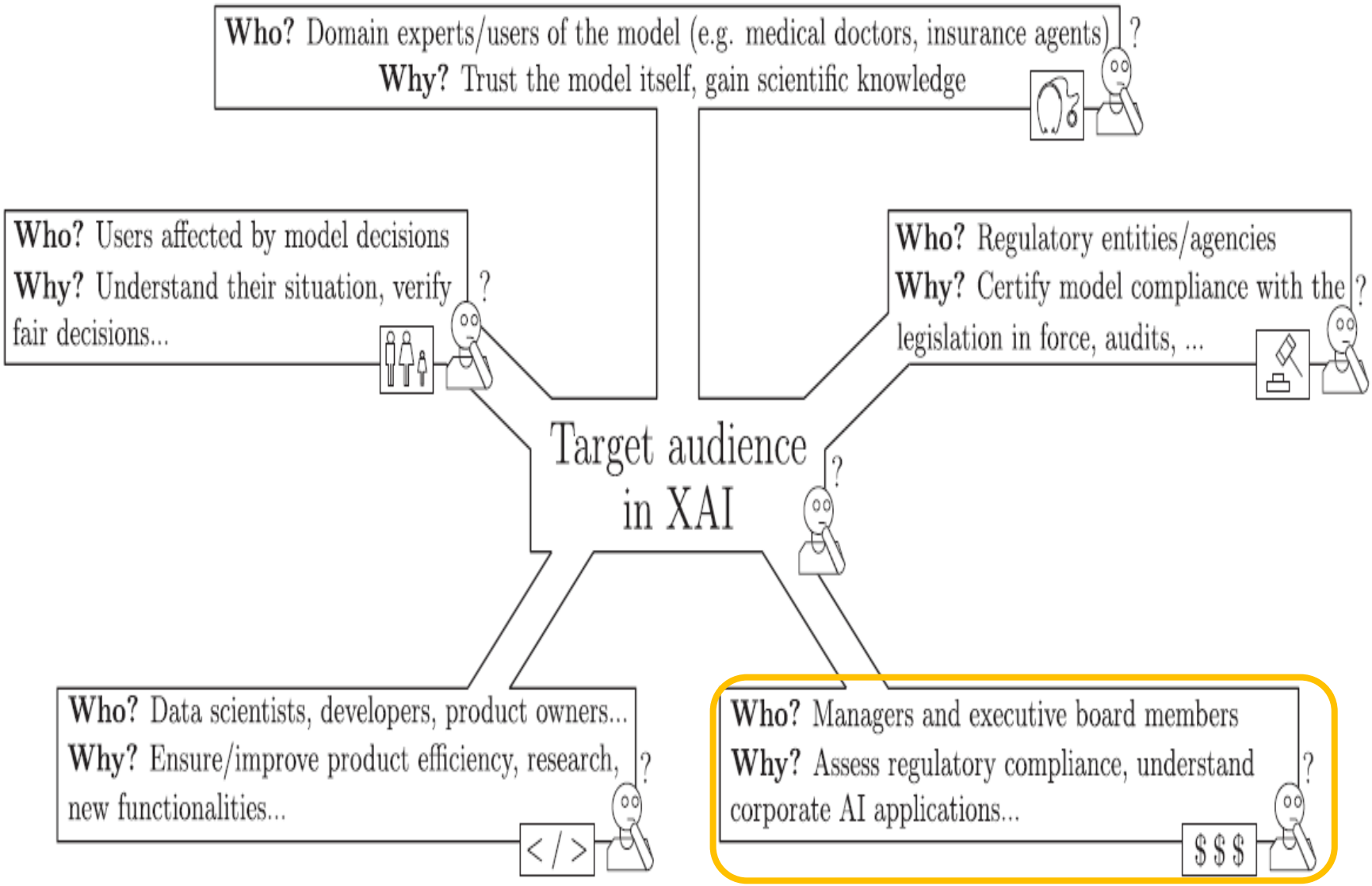
- Explicabilidade precisa levar em considerar a audiência

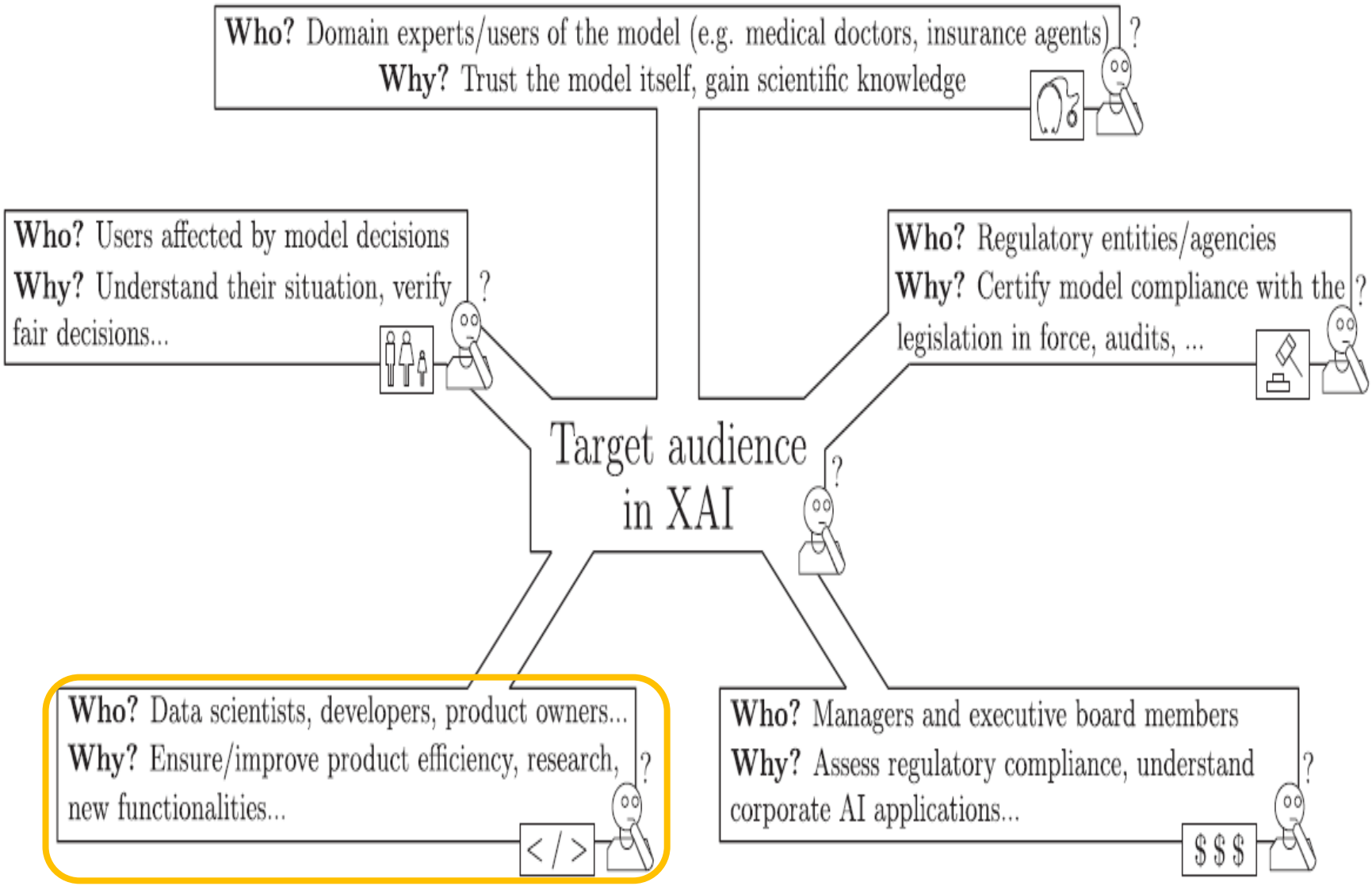
"Dado um **determinado público**, explicabilidade refere-se aos detalhes e razões fornecidos por um modelo para tornar seu funcionamento claro ou fácil de entender"

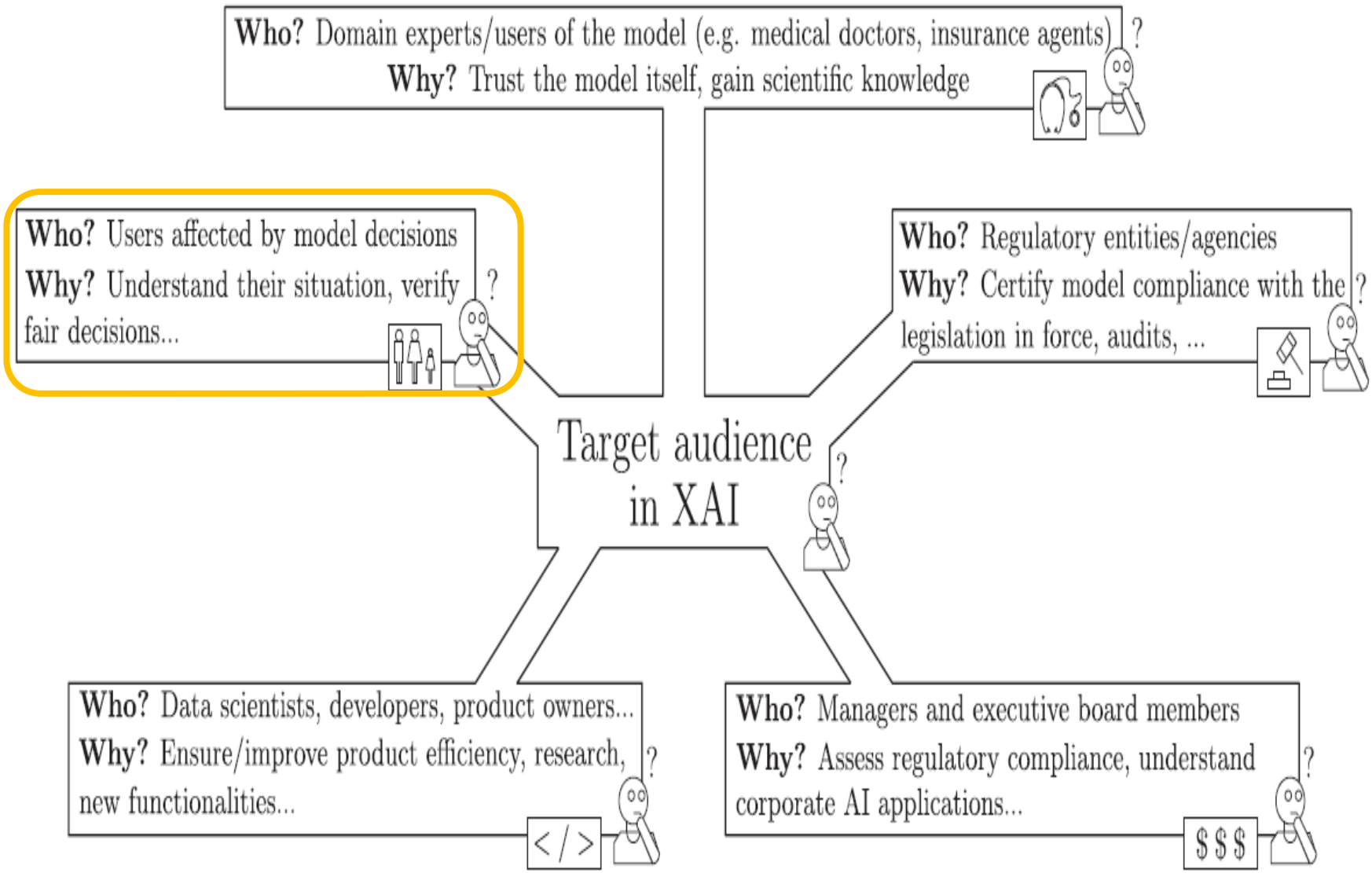

Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge
Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...
Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...
Target audience in XAI
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...
Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...
$ $ $


Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge
Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...
Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...
Target audience in XAI
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...
Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...
$ $ $

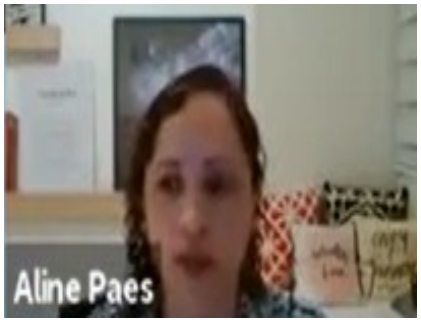

Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge
Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...
Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...
Target audience in XAI
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...
Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...
$ $ $


Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge
Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...
Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...
Target audience in XAI
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...
Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...
$ $ $


Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge
Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...
Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...
Target audience in XAI
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...
Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...
$ $ $

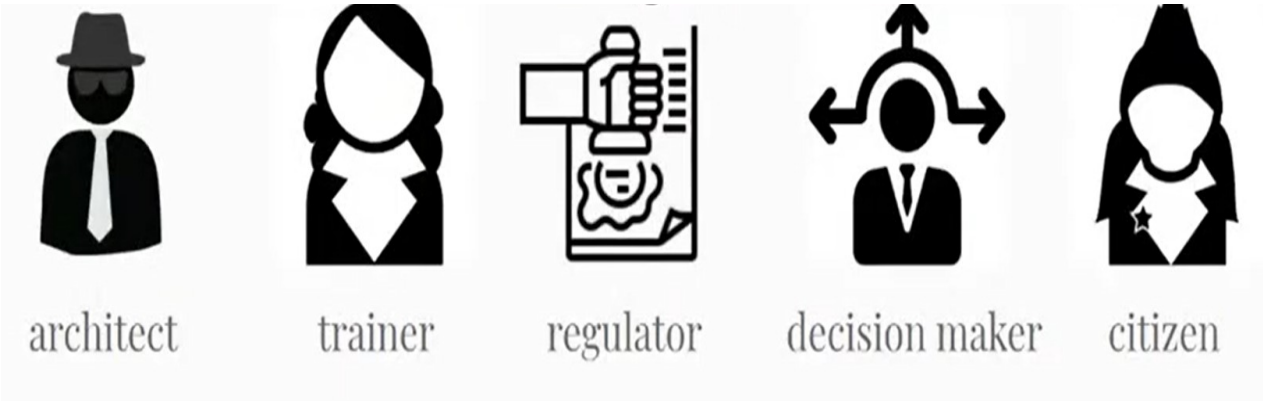

Who? Domain experts/users of the model (e.g. medical doctors, insurance agents)
Why? Trust the model itself, gain scientific knowledge
Who? Users affected by model decisions
Why? Understand their situation, verify fair decisions...
Who? Regulatory entities/agencies
Why? Certify model compliance with the legislation in force, audits, ...
Target audience in XAI
Who? Data scientists, developers, product owners...
Why? Ensure/improve product efficiency, research, new functionalities...
Who? Managers and executive board members
Why? Assess regulatory compliance, understand corporate AI applications...
$ $ $

Dra. Alina Paes em sua Palestra: “A Arte de Fazer Entender: Explainable AI para Impacto Social Positivo” (2021)
https://www.youtube.com/watch?v=z1iyYHpjcvs&t=3147s



1° Ponto de Vista: Para quem estou dando a explicação?

Depende muito da pessoa para qual aquela explicação esta sendo fornecida.

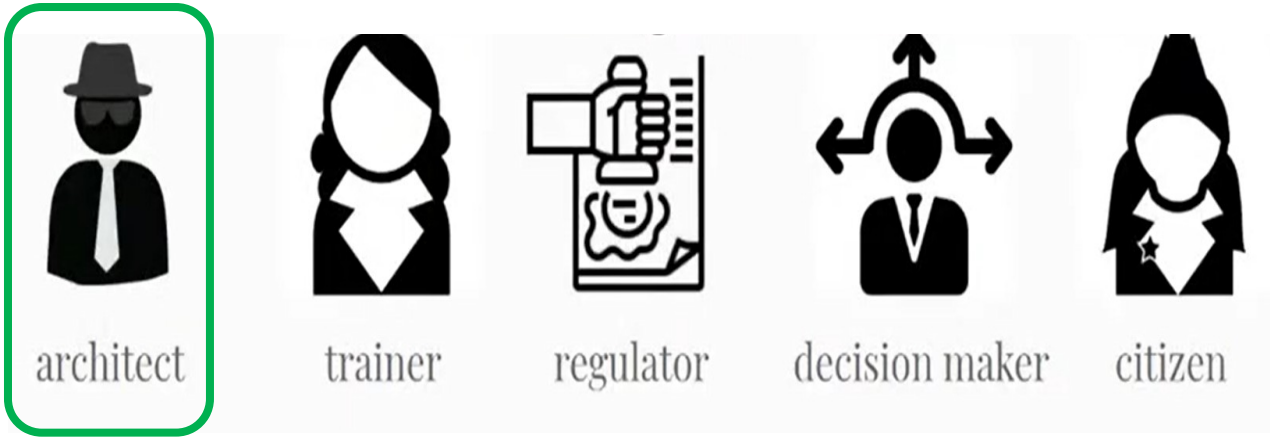


Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Depende muito da pessoa para qual aquela explicação esta sendo fornecida.

Pessoa que criou o programa que vai ser treinado, quem conhece a arquitetura da solução.

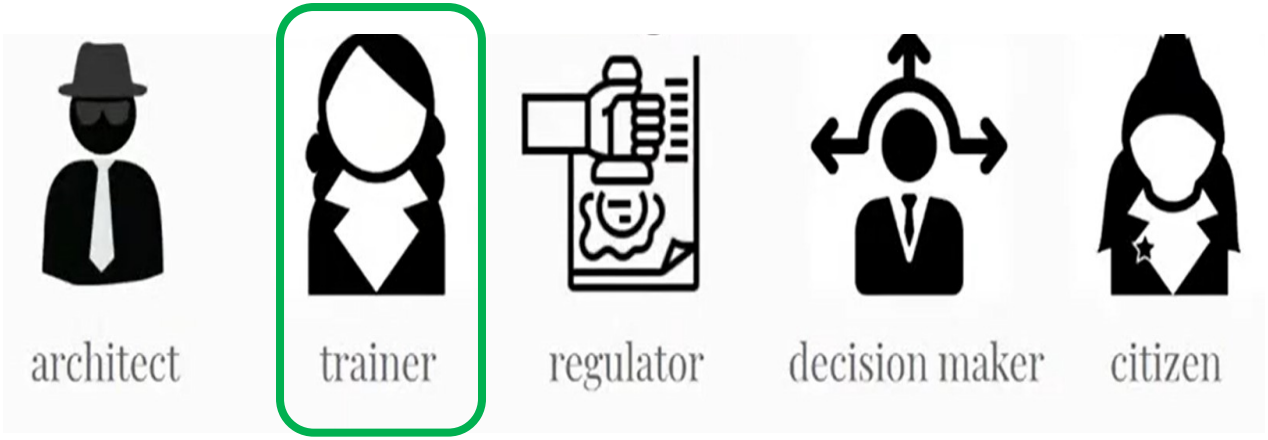


Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Depende muito da pessoa para qual aquela explicação esta sendo fornecida.

Pessoa vai pegar essa arquitetura, esta usando a solução para treinar ou resolver um problema a partir de um conjunto de dados.

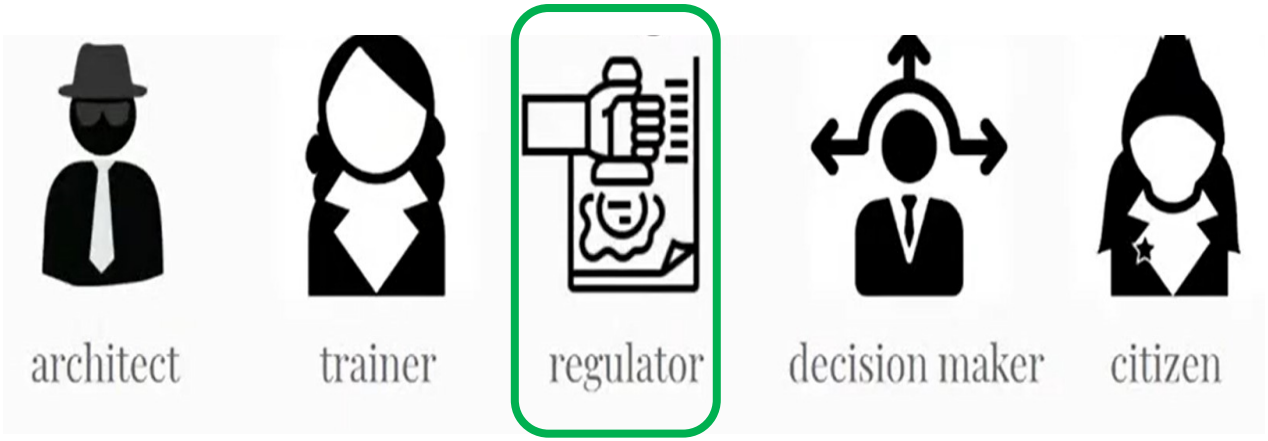


Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Depende muito da pessoa para qual aquela explicação esta sendo fornecida.

Quer saber se o sistema não esta fazendo nada de ilegal ou errado.

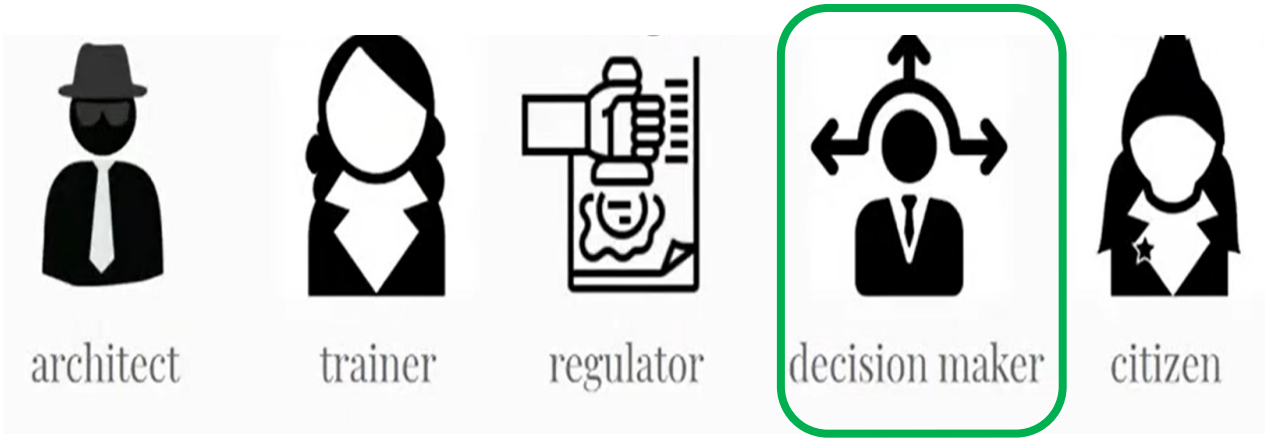


Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Depende muito da pessoa para qual aquela explicação esta sendo fornecida.

Pessoa que está recebendo o modelo e decidindo alguma coisa com base nesse modelo.

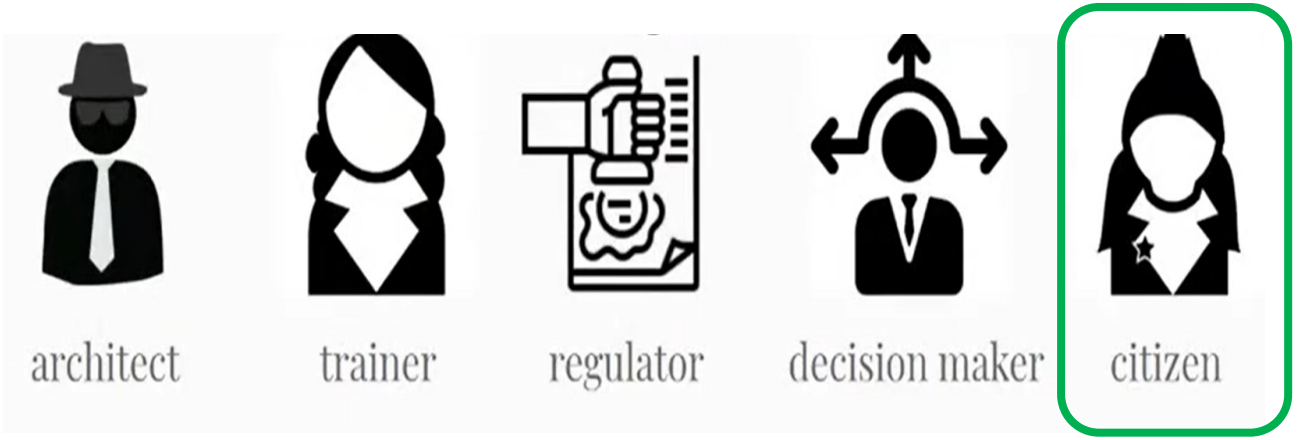


Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Depende muito da pessoa para qual aquela explicação esta sendo fornecida.

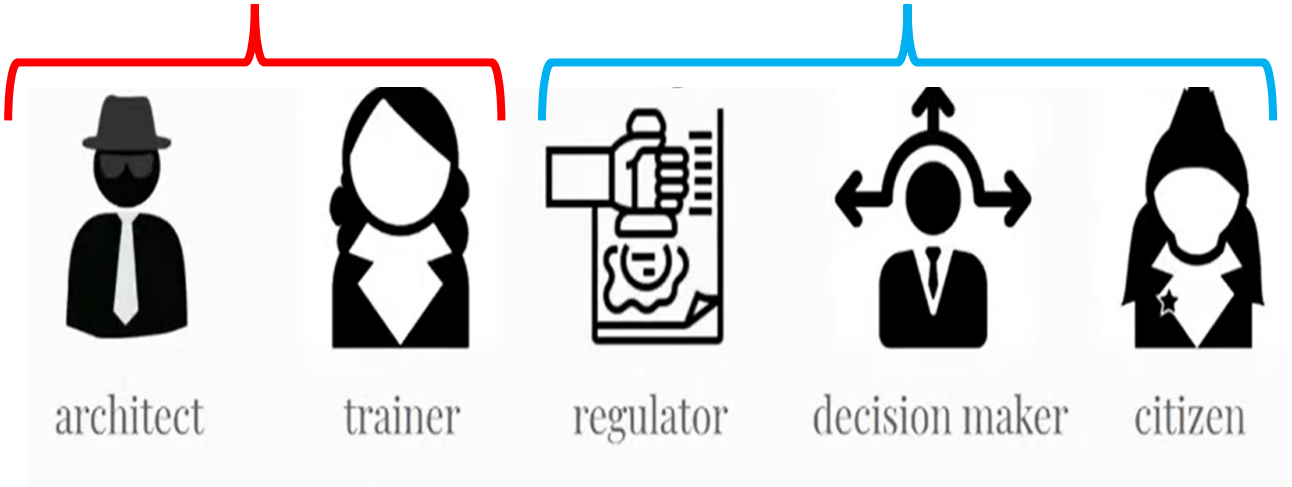
Usuário final cuja vida pode estar sendo implicada a partir de aquela decisão.



Terminologia adaptada de: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

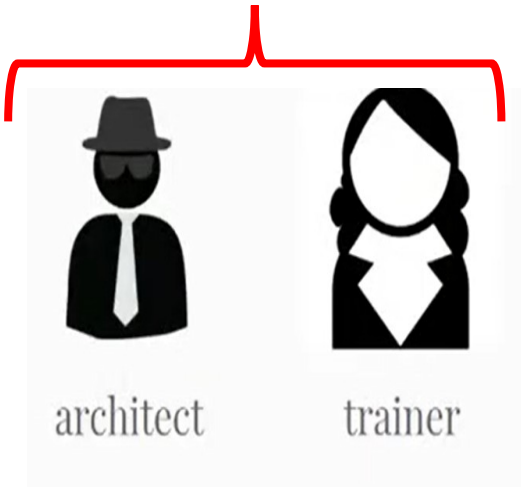
Dependendo de para quem você quer dar uma resposta podemos ter **<u>diferentes formas de produzir essa resposta</u>**, porque cada um deles vai estar interessado em um aspecto diferente:



Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

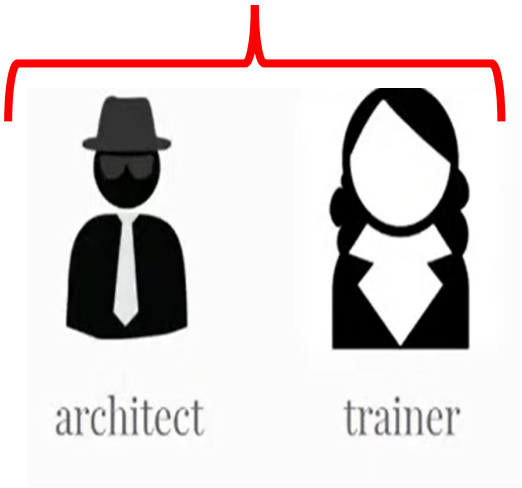
Que tipo de explicação?



Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Que tipo de explicação?

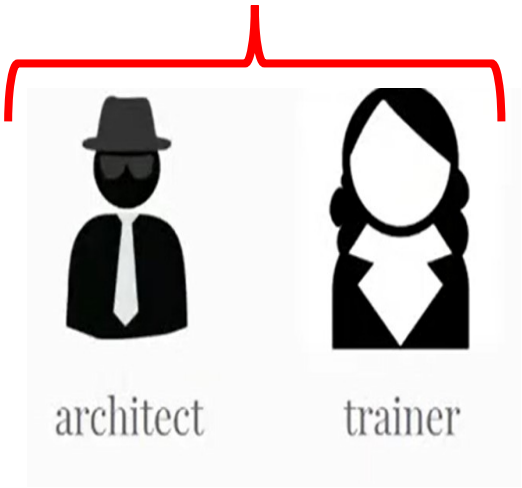


Estão interessados por que que o modelo chegou numa certa resposta e não em outra? O que está fazendo o modelo internamente? Ponto de vista do funcionamento interno.

Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Que tipo de explicação?



Estão interessados por que que o modelo chegou numa certa resposta e não em outra? O que está fazendo o modelo internamente? Ponto de vista do funcionamento interno.

Ferramentas de Interpretação

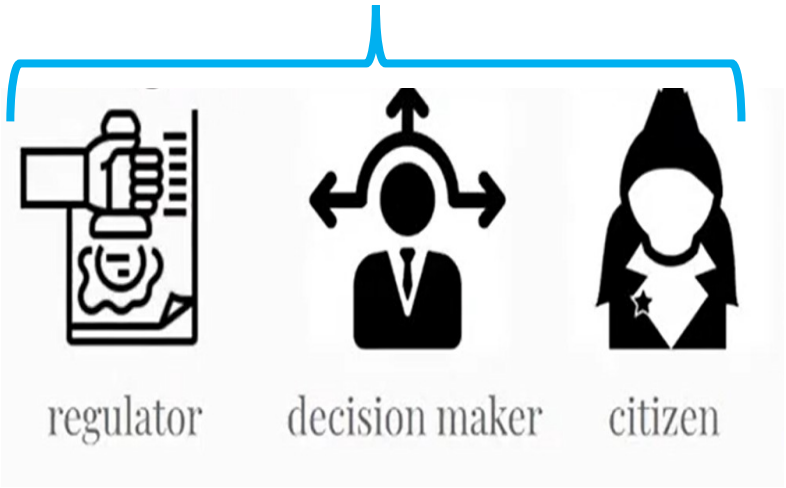
Mostram detalhes sobre a arquitetura. Ex:

- https://pair-code.github.io/lit/

Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Que tipo de explicação?

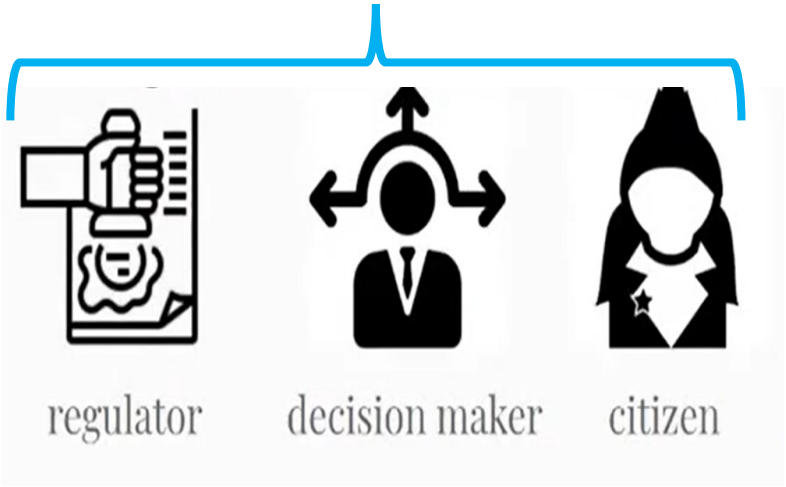


Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Tem que dar explicações que sejam de fato mais interpretáveis.

Que tipo de explicação?

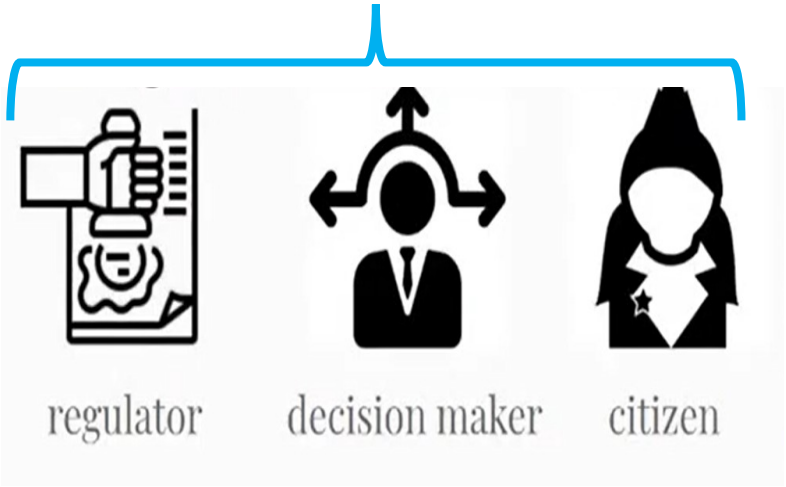


Terminologia adaptada de:
https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

1° Ponto de Vista: Para quem estou dando a explicação?

Tem que dar explicações que sejam de fato mais interpretáveis.

Que tipo de explicação?



Ferramentas de Interpretação

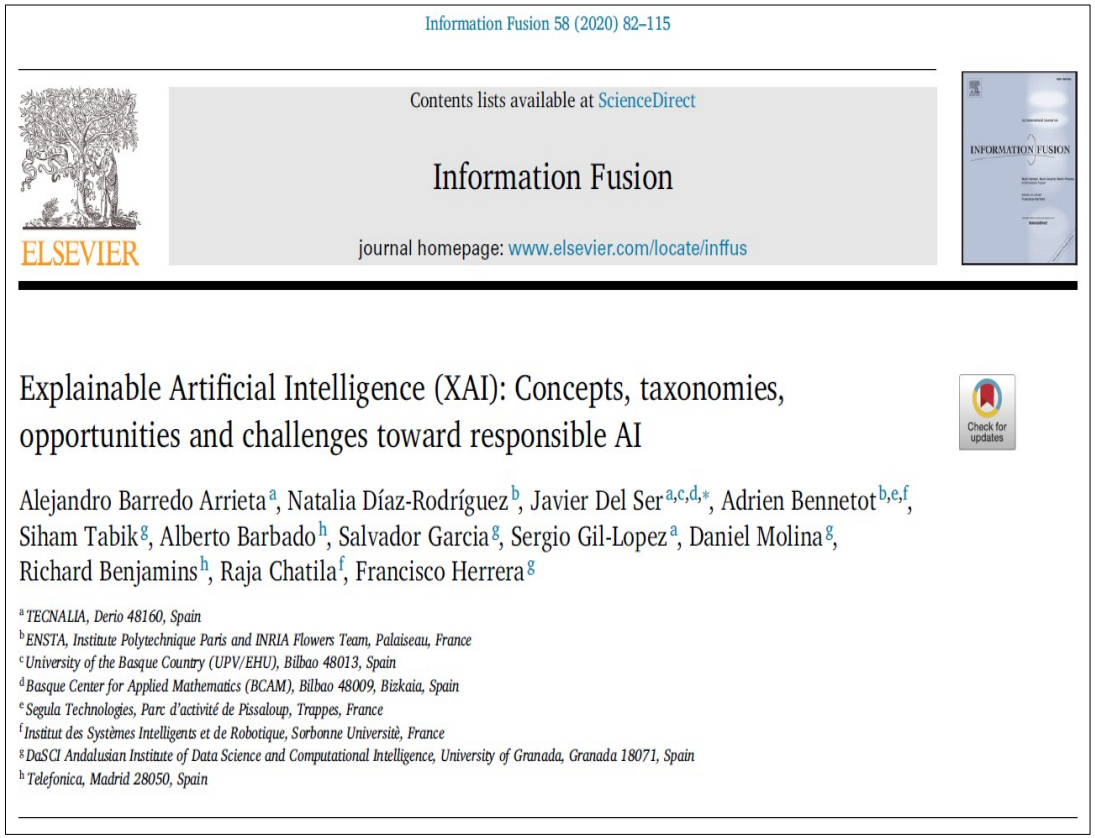
São utilizadas representações que sejam semanticamente possíveis de ser entendidas.

- Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. https://arxiv.org/pdf/1905.00563.pdf
- Pezeshkpour, P., Tian, Y., & Singh, S. (2019). Investigating robustness and interpretability of link prediction via adversarial modifications. https://arxiv.org/pdf/2010.00711.pdf

Terminologia adaptada de: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8017583

Como conseguir explicabilidade em IA?

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." Information fusion 58 (2020): 82-115 https://www.sciencedirect.com/science/article/pii/S1566253519308103?via%3Dihub



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]

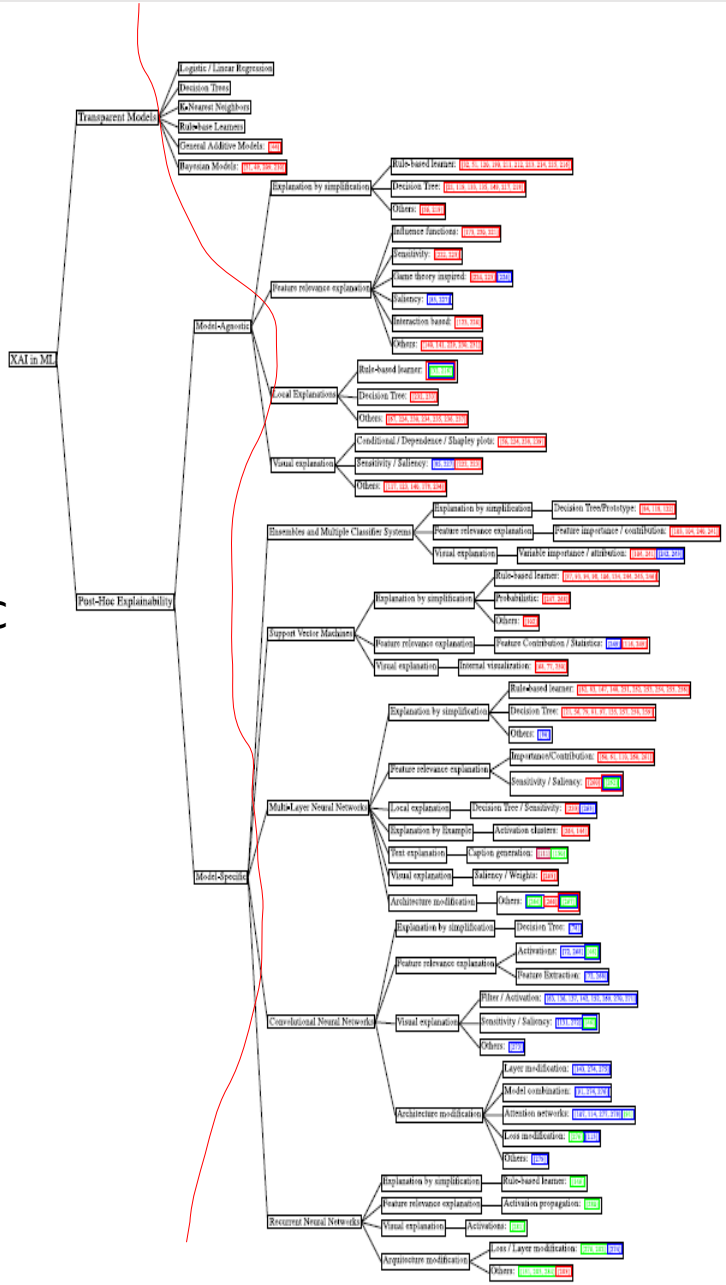
[a] TECNALIA, Derio 48160, Spain
[b] ENSTA, Institute Polytechnique Paris and INRIA Flowers Team, Palaiseau, France
[c] University of the Basque Country (UPV/EHU), Bilbao 48013, Spain
[d] Basque Center for Applied Mathematics (BCAM), Bilbao 48009, Bizkaia, Spain
[e] Segula Technologies, Parc d'activité 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

Como conseguir explicabilidade em IA?

- Modelos* Transparentes
- Explicabilidade Post-Hoc (tem que analisar o modelo)
  - Métodos agnósticos (independentes de modelo)
  - Métodos modelo-específicos

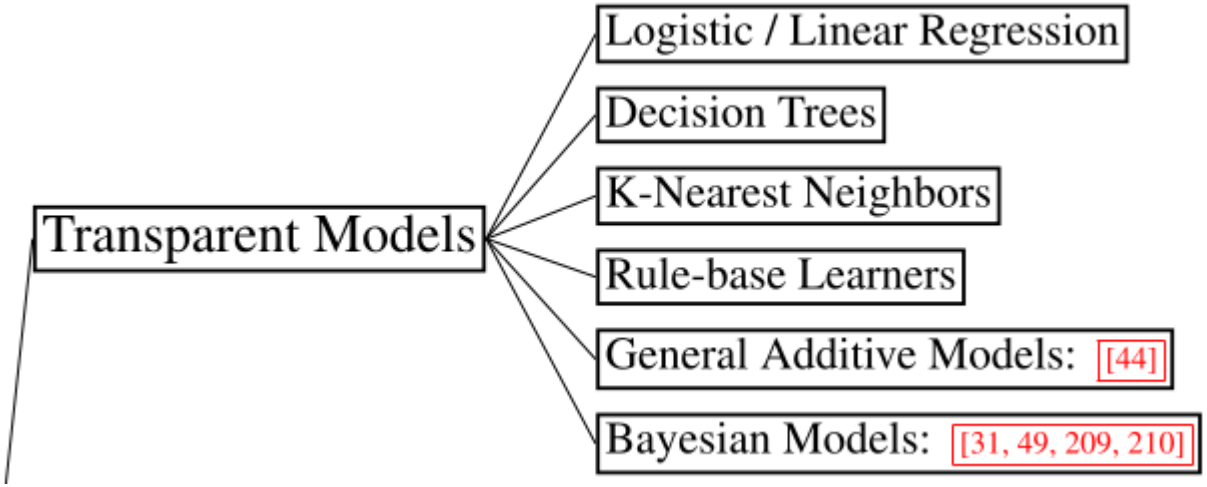
* nesta parte chamamos de modelos os algoritmos (classificação/regressão)

- Modelos* Transparentes
- Explicabilidade Post-Hoc
  - Métodos agnósticos
  - Métodos modelo-específicos



Métodos de explicabilidade

Modelos Transparentes

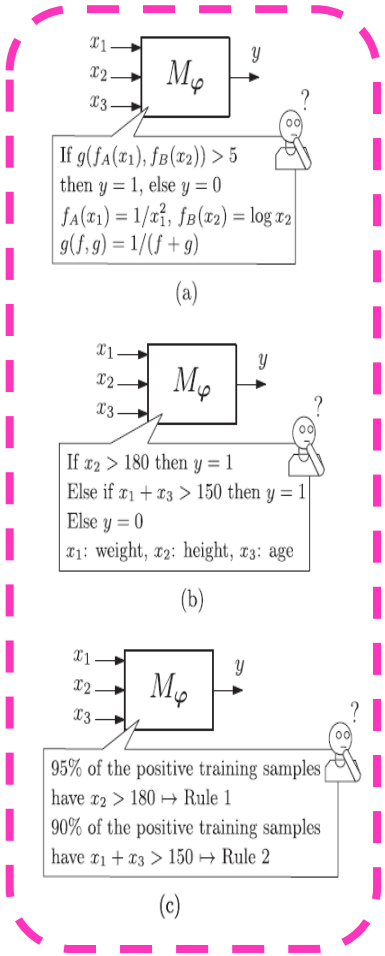


Níveis de Transparência

1. Simulabilidade

2. Decomponibilidade

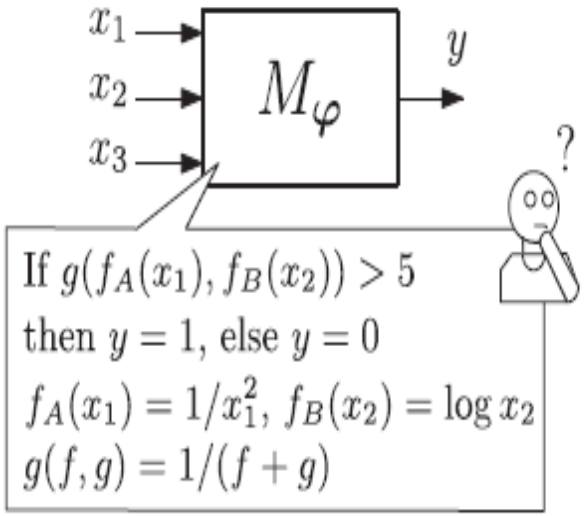
3. Transparência algorítmica



Níveis de Transparência

1. Simulabilidade

A capacidade de um modelo ser **simulado**, ou seja requer que o modelo seja **autocontido** o suficiente para que um humano pense e raciocine sobre ele como um **todo**.

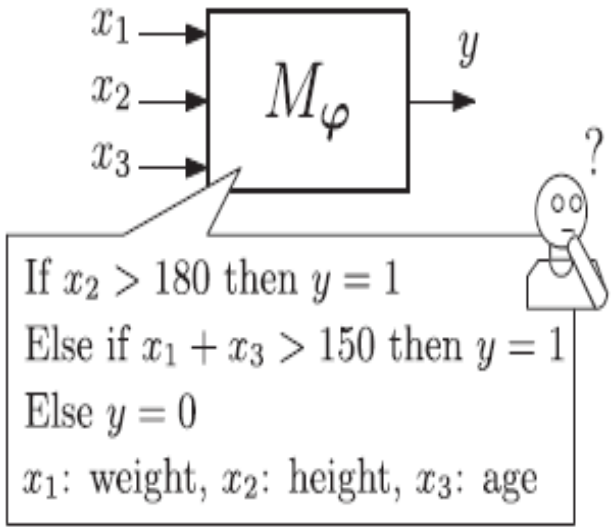


(a)

Níveis de Transparência

2. Decomponibilidade

(Decomposto) A capacidade de explicar cada uma das **partes** do modelo (entrada, parâmetros e cálculo). A Decomponibilidade requer que cada **entrada** seja prontamente interpretável (características do PCA não são). A restrição adicional é que todas as partes do modelo devem ser compreensíveis por um ser humano **sem a necessidade de ferramentas adicionais**.

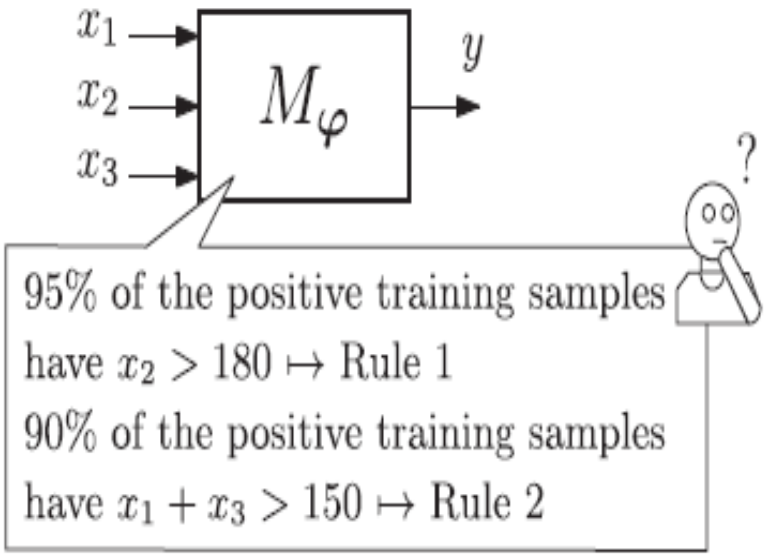


(b)

Níveis de Transparência

3. Transparência Algorítmica

Lida com a capacidade do usuário de entender o processo do modelo para produzir qualquer saída a partir de seus dados de entrada. A principal restrição é que o modelo deve ser totalmente explorável por meio de análises e métodos matemáticos.



(c)

Table 2
Overall picture of the classification of ML models attending to their level of explainability.

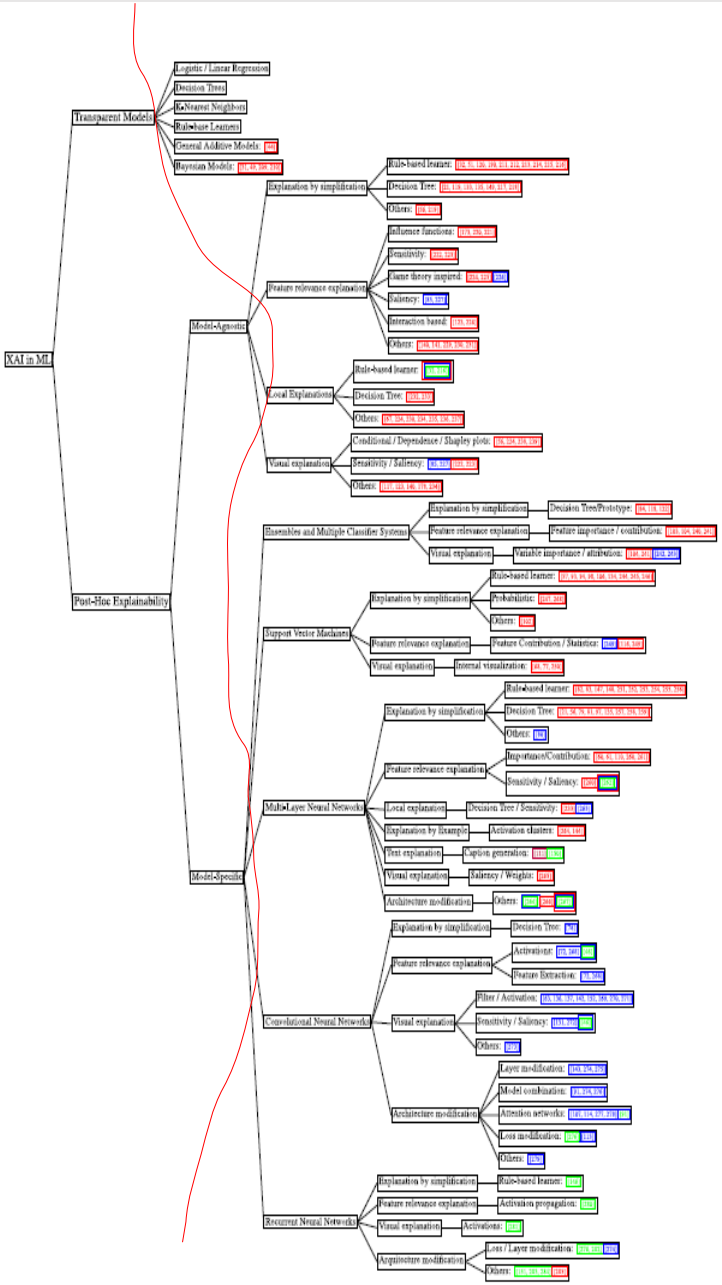
Model	Transparent ML Models			Post-hoc analysis
	Simulatability	Decomposability	Algorithmic Transparency	
Linear/Logistic Regression	Predictors are human readable and interactions among them are kept to a minimum	Variables are still readable, but the number of interactions and predictors involved in 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

Zoom

Exemplo: Árvores de Decisão

Modelo	Simulabilidade	Decomponibilidade	Transparência Algorítmica	Análise Post-Hoc
Árvores de Decisão	Um ser humano pode simular e obter a previsão de um árvore de decisão por conta própria, sem exigir nenhum base matemática.	O modelo contém regras que não alteram os dados por qualquer coisa, e preserva sua legibilidade.	Regras legíveis por humanos que explicar o conhecimento aprendido com os dados e permite para uma compreensão direta do processo de previsão.	X

- Modelos* transparentes
- Explicabilidade post-hoc
  - Métodos agnósticos
  - Métodos modelo-específicos

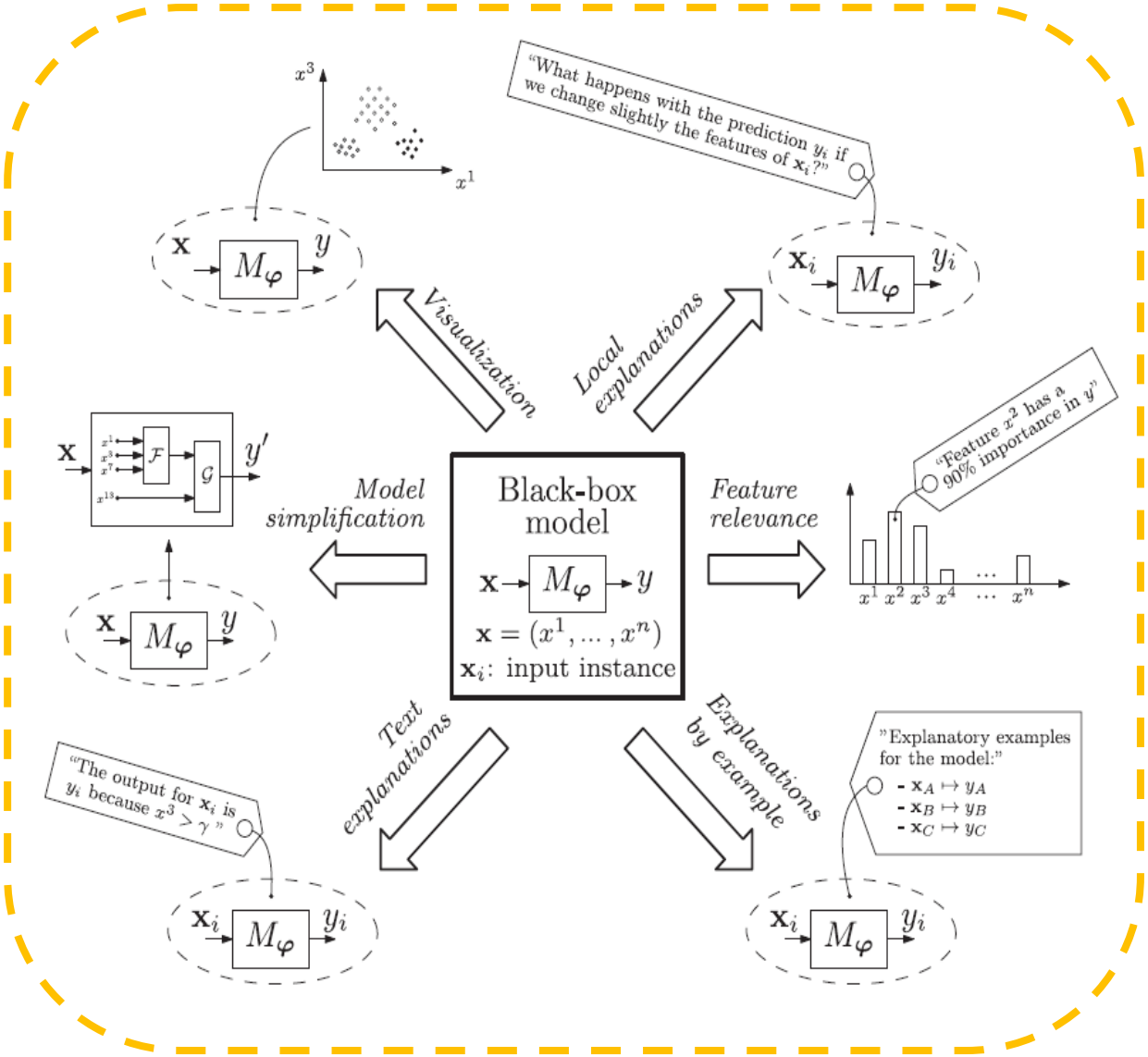


Métodos de explicabilidade

Explicabilidade post-hoc

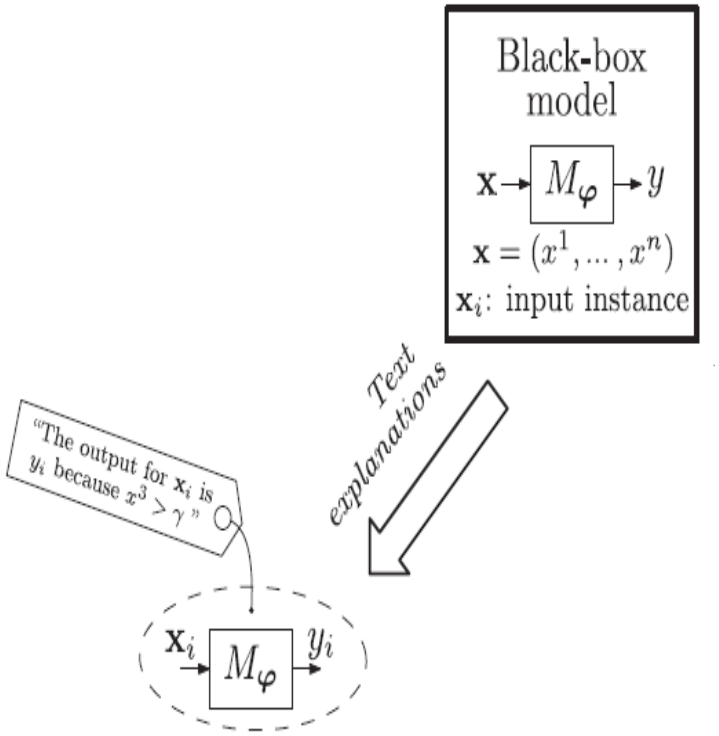
Para modelos que não são facilmente interpretáveis por *design*, recorrendo diversos meios para melhorar sua interpretabilidade.

Abordagens de explicabilidade post-hoc



1. Explicações de Texto:

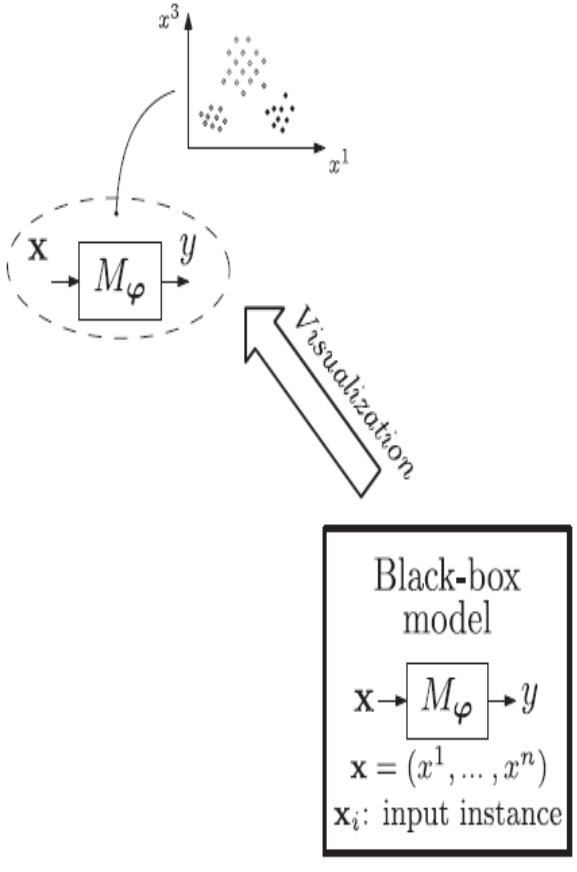
Lidam com o problema de trazer explicabilidade para um modelo por meio de aprender a gerar explicações de **texto** que ajudem a explicar os resultados do modelo.



As explicações de texto também incluem todos os métodos que geram símbolos que representam o funcionamento do modelo, esses símbolos podem retratar a lógica do algoritmo.

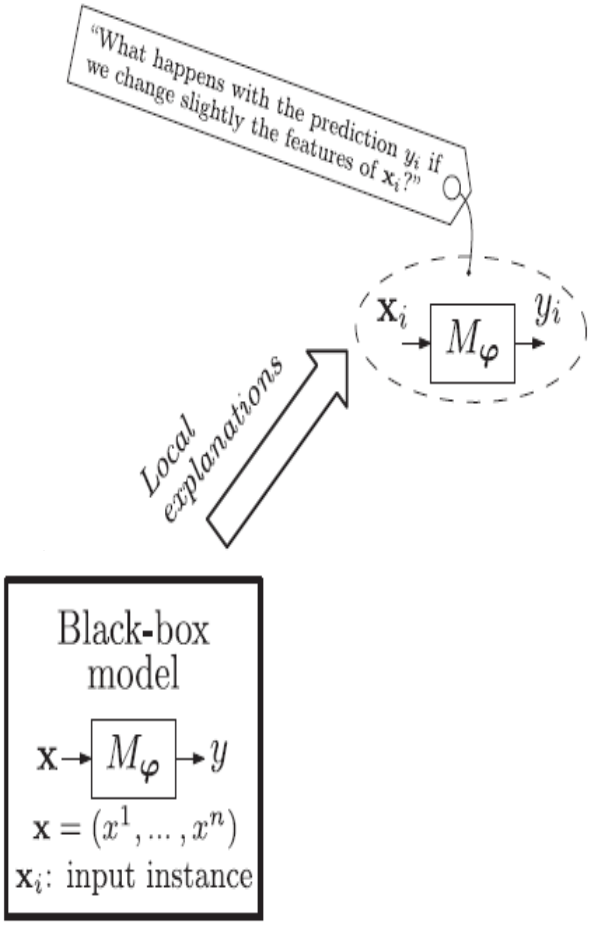
2. As Explicações Visuais:

Visam a visualizar o comportamento do modelo. Muitos deles vêm acompanhados de técnicas de redução de dimensionalidade que permitem uma visualização humana interpretável. São consideradas a forma mais adequada de introduzir interações complexas dentro das variáveis envolvidas no modelo.



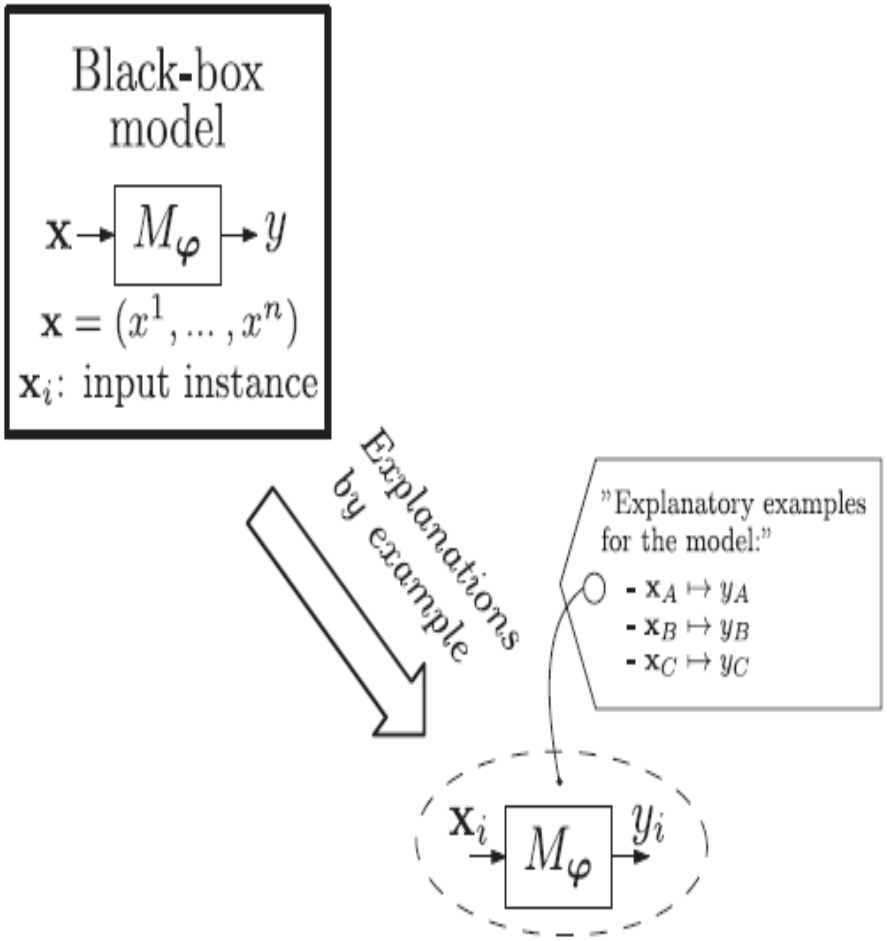
3. As Explicações Locais:

Segmentam o espaço de solução e dando explicações a subespaços de solução menos complexos que são relevantes para todo o modelo.



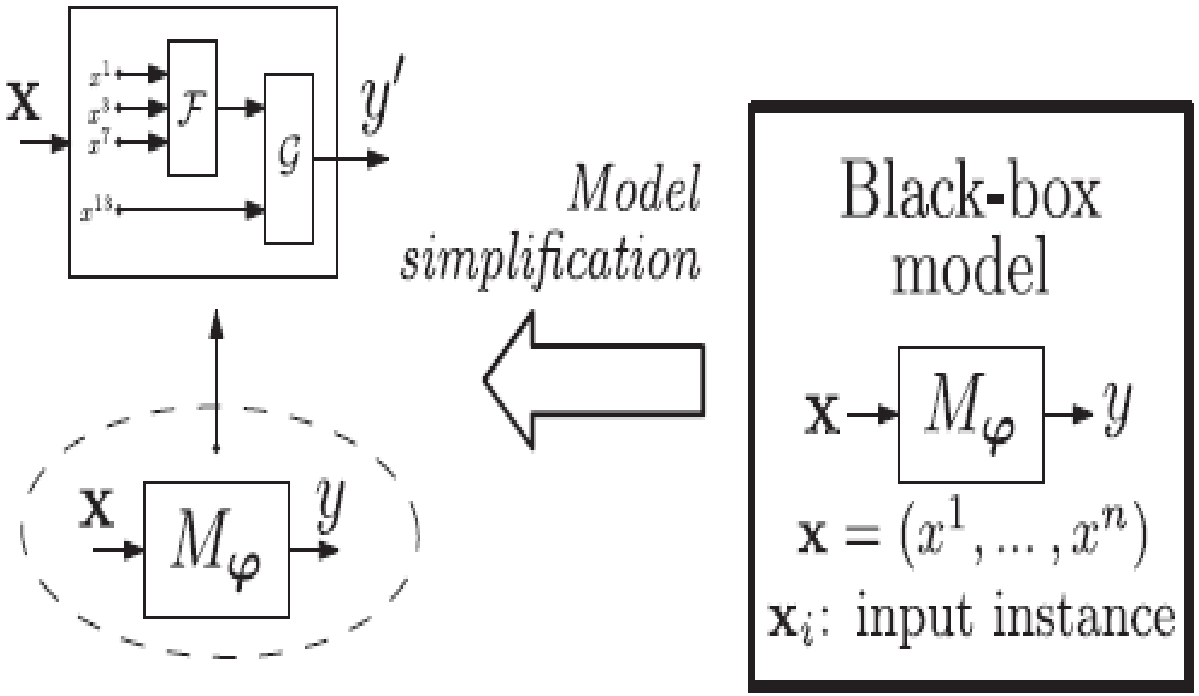
4. As Explicações por Exemplo:

Estão centradas principalmente na extração de **exemplos representativos** que aprendam as relações e correlações internas encontradas pelo modelo em análise.



5. As Explicações por Simplificação:

São técnicas onde todo um novo sistema é reconstruído com base no modelo a ser explicado. Esse novo modelo simplificado tenta otimizar sua semelhança com seu funcionamento anterior, reduzindo sua complexidade e mantendo uma pontuação de desempenho semelhante.



5. As Explicações por Simplificação:

Ex: **LIME** (Local Interpretable Model-agnostic Explanations), útil quando se deseja entender como um modelo específico chegou a uma determinada previsão para uma **instância individual**.

Processo geral:

1. Seleção de Instância para a qual se quer explicar a previsão do modelo

2. Gerar várias versões perturbadas da instância selecionada, introduzindo pequenas alterações ou ruído nos dados, preservando a maioria das características dos dados originais.

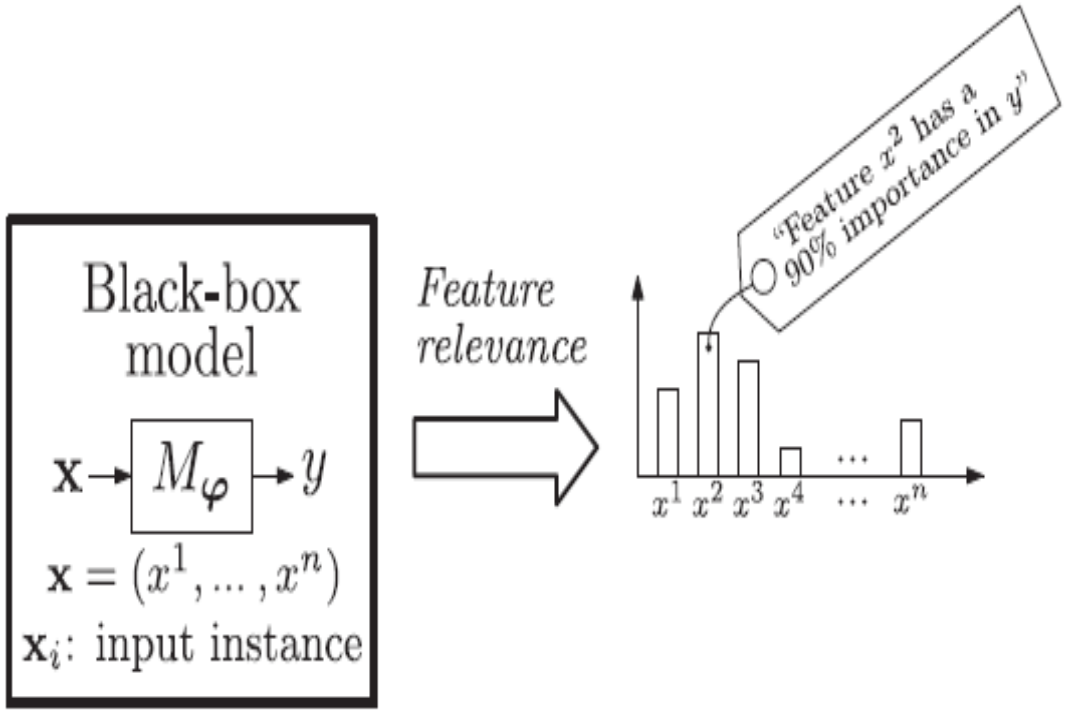
3. Para cada versão perturbada da instância, calcular as previsões usando o modelo complexo. Isso cria um conjunto de dados onde as predições do modelo complexo são usadas como variáveis-resposta.

4. Treinar um modelo simples e interpretável (como uma regressão linear ou árvore de decisão) para ajustar-se aos dados gerados.

5. Utilizar o modelo simples treinado para atribuir importância ou pesos às diferentes características, indicando como cada uma contribui para a previsão final do modelo complexo para essa instância específica.

6. As Explicações por Relevância de Características:

Computam uma pontuação de relevância para suas variáveis, essas pontuações quantificam a afeição que um recurso tem sobre a saída do modelo. A comparação das pontuações entre as diferentes variáveis revela a importância que o modelo atribui a cada uma dessas variáveis na produção de sua saída.



6. As Explicações por Relevância de Características:

Exemplos: *feature importances* em *Random Forests*, e SHAP

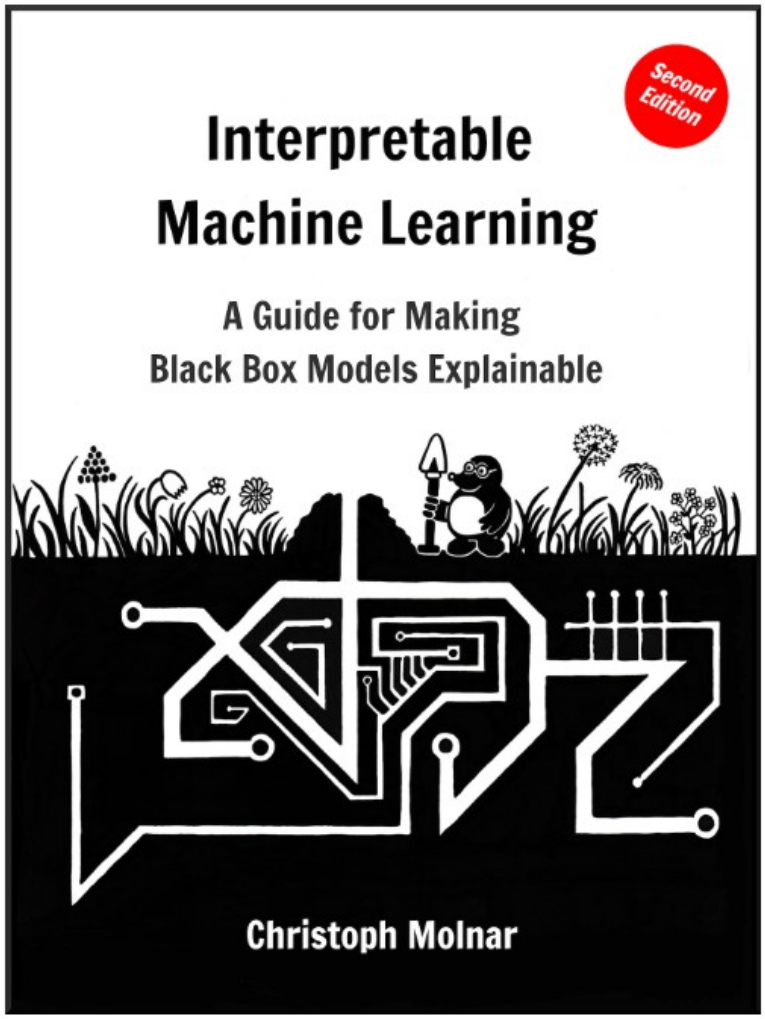
SHAP (SHapley Additive exPlanations) se baseia na teoria dos jogos para calcular a contribuição de cada variável:

1. Seleciona-se um conjunto de dados de referência (para explicar). Pode ser um ponto de dados específico ou um conjunto médio de características.

2. Calcula a contribuição de cada característica para a diferença entre a predição do modelo e um *baseline*. Ele analisa todas as possíveis combinações de características para calcular essas contribuições de forma consistente.

3. Com base nos resultados, gera um valor de importância para cada característica, indicando como cada ums influencia a mudança na predição do modelo, seja aumentando ou diminuindo o valor previsto.

Além de oferecer uma explicação local para uma única predição, o SHAP também pode fornecer uma visão geral, mostrando a importância relativa de cada recurso ao longo de todo o conjunto de dados.



https://christophm.github.io/interpretable-ml-book/