Inteligência Artificial Interpretável ou Explicável?

Prof. Denis Martins

DCM | FFCLRP | USP

Contextualizando

Complexidade

Modelos de modernos de IA possuem bilhões de parâmetros que controlam seus comportamentos.

Compreensão

Necessidade de entender comportamento e processos de decisão de modelos de IA.

Equidade e Transparência

Garantia de justiça algorítmica, identificação de vieses e investigação de problemas de dados.

> Baseado em: ARRIETA, Alejandro Barredo et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, v. 58, p. 82-115, 2020.

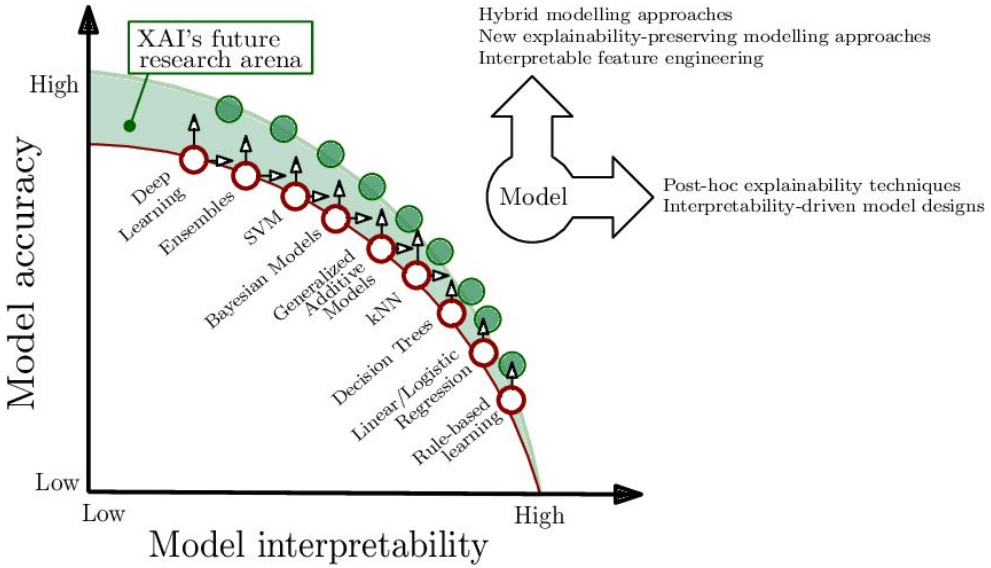
Modelos Interpretáveis

Simplicidade

Mecanismos de decisão matematicamente simples, compreensíveis por um ser humano sem a necessidade de ferramentas externas.

Regras Claras

O relacionamento entre as variáveis de entrada e a saída é baseado em regras transparentes e perfeitamente rastreáveis.



Baseado em: ARRIETA, Alejandro Barredo et al. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information fusion, v. 58, p. 82-115, 2020.

Interpretação de Modelos

Post-Hoc

Aplicar técnicas externas para aproximar ou explicitar o raciocínio de modelos complexos.

Variação

Se eu mudar ligeiramente esta entrada, como a saída muda?
Quais partes da entrada foram mais importantes neste ponto específico?

SHAP

Baseado em: CHRISTOPH, Molnar. Interpretable machine learning: A guide for making black box models explainable. 2020.

Compreender o que um modelo fez ou pode ter feito.

Compreender a estrutura interna do modelo e como suas variáveis de entrada afetam diretamente as saídas.

Propriedade intrínseca ao design do algoritmo.

Interpretabilidade

≠

Explicabilidade

Compreender os motivos do comportamento do modelo (insights sobre as causas de suas decisões).

Permitir a auditoria do raciocínio.

Gerar uma justificativa compreensível para um resultado específico, mesmo que o modelo subjacente seja complexo.

Baseado em: GILPIN, Leilani H. et al. Explaining explanations: An overview of interpretability of machine learning. In: 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018. p. 80-89.

Explicação para quem?

Uma explicação eficaz deve ser um **produto customizado**, cujo nível de detalhe e foco narrativo são determinados pela capacidade cognitiva, interesse e poder de ação do receptor.

Evitar dois erros:

Sobrecarga Cognitiva

Cognitive Overload

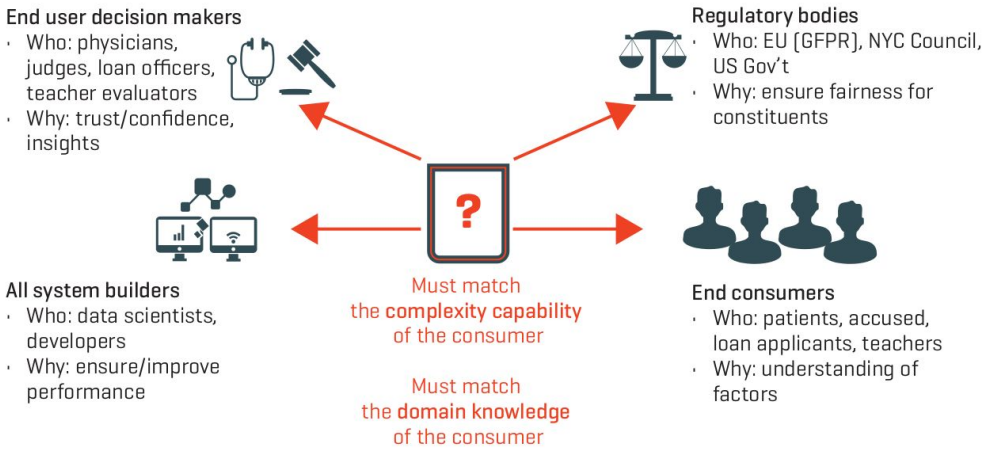
Apresentar jargões técnicos demais para um leigo (excesso de detalhes matemáticos).

Simplificação Excessiva

Oversimplification

Ocultar complexidades críticas sob uma narrativa muito simples, levando a decisões errôneas.

Figure 1. The many groups interested in explainable AI.



Baseado em: HIND, Michael. Explaining explainable AI. XRDS: Crossroads, The ACM Magazine for Students, v. 25, n. 3, p. 16-19, 2019..

Inteligência Artificial é, antes de tudo, Software

Produção carregada de visão de mundo, cultura, propósito, viés...

Sugeita a planejamento, governança, regulação, versionamento...

Aspectos Sociotécnicos

Desenvolver software é projetar simultaneamente uma tecnologia e uma linguagem (de comunicação).

A **Semiotic Ladder** de Stamper nos permite analisar sistemas de informação através de seis níveis interdependentes.

O sucesso real de qualquer sistema de software depende do alinhamento entre a infraestrutura técnica (base) e as funções de uso humano (topo).

Leia mais em:
https://tagg.org/rmt/Writings/StamperSemioticLadder.htm

Funções Humanas

6. MUNDO SOCIAL
Valores, crenças, compromissos, cultura, contratos e leis

5. PRAGMÁTICO
Intenções, comunicação, conversações e negociações

4. SEMÂNTICO
Significados, proposições, validade e verdade

Plataforma de TI

3. SINTÁTICO
Estruturas formais, dados, linguagem, lógica e software

2. EMPÍRICO
padrões, ruído, redundância, capacidade e eficiência

1. FÍSICO
Hardware, redes, dispositivos e meios materiais

Muito Obrigado!