

# **From human explanations to explainable AI: Insights from constrained optimization**

Inteligência Artificial Simbólica para Ciência de Dados

Grupo 15

*David Franco, nº 110733*

*João Dias, nº 110305*

2024/2025

# INTRODUÇÃO

# CONCEITOS IMPORTANTES

01

## **EXPLAINABLE AI (XAI)**

02

**#1** As redes neurais destacam-se pela sua capacidade de resolver problemas complexos, mas a sua natureza opaca e a falta de interpretabilidade limitam a aceitação em sistemas críticos, onde é essencial que as decisões sejam compreendidas e justificadas (lbs et al., 2024).

03

04

05

**#2** A Inteligência Artificial Explicável (XAI) surge como uma abordagem para mitigar esta falta de transparência, permitindo justificar e compreender decisões, o que promove a confiança em determinados contextos (Bassan & Katz, 2023).

06

07

**#3** Domínios como a saúde e o planeamento energético exigem explicações claras, dado o impacto significativo que as decisões automatizadas podem ter em indivíduos e populações (Hoffman et al., 2018).

# PROBLEMAS DE OTIMIZAÇÃO COM RESTRIÇÕES

01

## COMPLEXIDADE INERENTE A PROBLEMAS REAIS

Problemas de otimização com restrições são amplamente aplicados em cenários reais mas têm muitos desafios associados quando existem múltiplas variáveis de decisão.



02

03

04

05

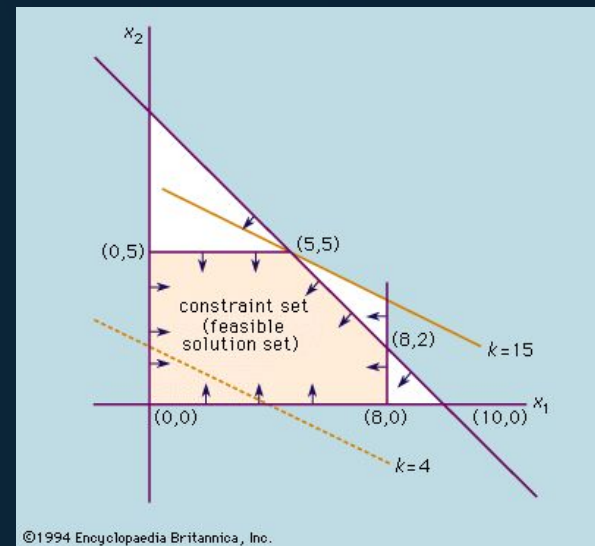
06

07

## EXEMPLO

Tome-se como exemplo a transição para energias renováveis, onde o objetivo é minimizar custos sem ultrapassar os limites de emissão (Ibs et al., 2024).

## REPRESENTAÇÃO GRÁFICA SIMPLES



©1994 Encyclopaedia Britannica, Inc.



# PROBLEMAS DE OTIMIZAÇÃO COM RESTRIÇÕES

01

02

03

04

05

06

07

Ferramentas como os **solvers** de programação linear têm sido fundamentais para encontrar **soluções matematicamente ótimas** em problemas complexos, garantindo eficiência computacional (lbs et al., 2024).

# PROBLEMAS DE OTIMIZAÇÃO COM RESTRIÇÕES

01

02

03

04

05

06

07

Embora ofereçam soluções ótimas, os solvers frequentemente geram **resultados em espaços de alta dimensionalidade**, o que dificulta a tradução desses resultados em explicações compreensíveis para utilizadores não técnicos (Ibs et al., 2024).

# EXPLAINABLE AI EM PROBLEMAS DE OTIMIZAÇÃO

01

02

03

04

05

06

07

**#1** Estratégias cognitivas humanas, como o uso de heurísticas, simplificam a resolução de problemas de otimização, proporcionando abordagens mais intuitivas que podem ser integradas em sistemas explicáveis (Longo et al., 2024).

**#2** Traduzir as soluções geradas por algoritmos em explicações claras e intuitivas é essencial para garantir que as decisões sejam compreensíveis e fiáveis, especialmente para os utilizadores finais (Ibs et al., 2024).

# FURNITURE FACTORY



# PARADIGMA DA FURNITURE FACTORY

O paradigma **Furniture Factory** é amplamente utilizado para o estudo de problemas de otimização e foi concebido para explorar como as pessoas resolvem problemas de otimização, investigando as suas **estratégias heurísticas** e a **complexidade cognitiva das explicações** apresentadas (Ibs et al., 2024).

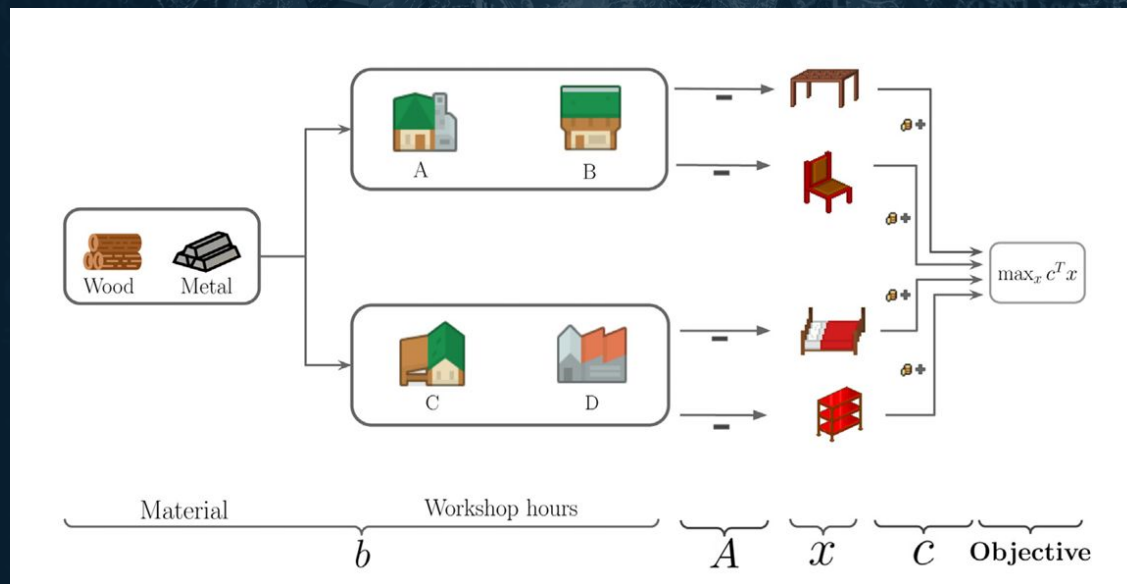


Fig. 1. Design estrutural do paradigma da Furniture Factory.

# **ESTUDO 1**

## **(CONCURRENT EXPLANATIONS)**

## ESTUDO 1 (CONCURRENT EXPLANATIONS)

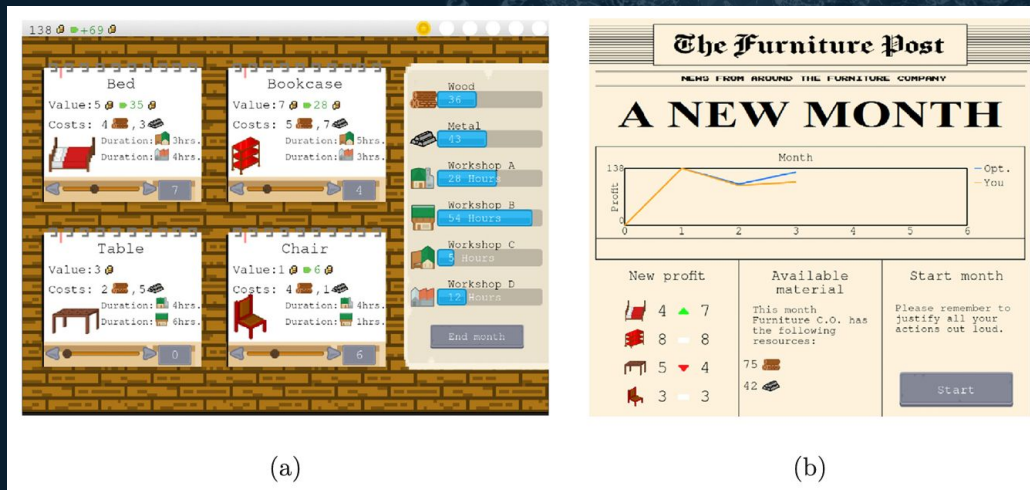


Fig. 2. Tarefa de exploração como jogo de computador.

(a) Interface principal com informações da tarefa e controlos. (b) Ecrã inicial com alterações no lucro e material disponível.

Recolha de dados qualitativos (explicações) e quantitativos (trajetórias de solução) para analisar como as pessoas resolvem problemas de otimização com restrições.

Utilização do conceito de backtracking:

- Os participantes ajustam os controladores para rever e corrigir escolhas
- Permite explorar e melhorar estratégias iterativamente, alcançando soluções mais eficientes

# PRINCIPAIS RESULTADOS

01  
02  
03  
04  
05  
06  
07

**Desempenho:** A performance média foi próxima do ótimo, com um valor médio de PUO de -2%, mas variou entre os participantes.

## Heurísticas Identificadas:

- **Balancing:** Estratégias para equilibrar recursos (ex: materiais ou tempo)
- **Profit-Oriented:** Foco no lucro imediato, priorizando itens mais lucrativos
- **Cost-Benefit:** Comparação entre custos e lucros para maximizar eficiência

## Análise das Comparações:

- Os participantes frequentemente compararam apenas dois itens de cada vez usando atributos como o lucro e custo dos materiais
- Comparações mais complexas foram feitas sequencialmente

**Foco Seletivo:** Os participantes priorizaram aspetos específicos (ex: custo dos materiais e lucro), raramente considerando todos os fatores ao mesmo tempo

$$\text{Performance (\%)}: \left( \frac{\text{participant score}}{\text{optimal score}} - 1 \right) \cdot 100.$$

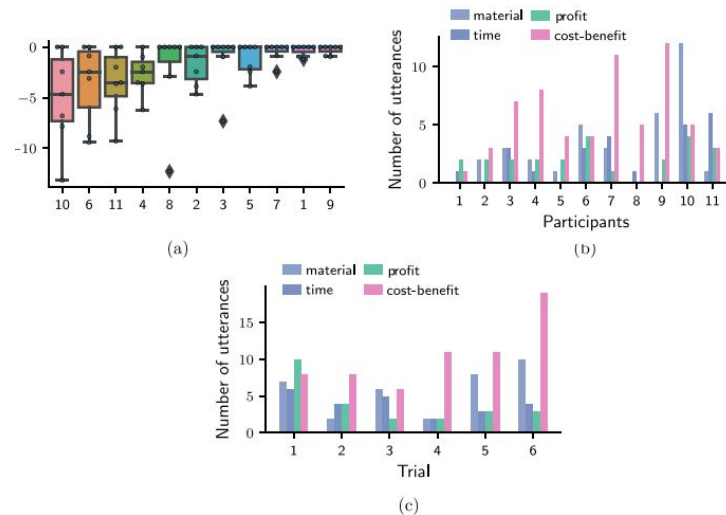


Fig. 3. Desempenho e frequência de estratégias.  
 (a) Pontuações Percentagem Abaixo do Ótimo por participante. (b) Número de estratégias mencionadas por base (lucro, etc.). (c) Estratégias usadas por tentativa.



## PRINCIPAIS CONCLUSÕES DO ESTUDO 1

- 01
- 02
- 03**
  - #1** Os participantes encontraram soluções eficientes, sugerindo que utilizam abordagens estruturadas
  - #2** Os participantes usaram estratégias baseadas no lucro, equilíbrio e custo-benefício
  - #3** As representações utilizadas eram simplificadas, focando-se em recursos específicos
  - #4** As estratégias e representações identificadas podem servir como base para a geração automática de explicações em sistemas de IA
- 04
- 05
- 06
- 07

# **ESTUDO 2 (POST-HOC EXPLANATIONS)**

## ESTUDO 2 (POST-HOC EXPLANATIONS)

01  
02  
03  
04  
05  
06  
07



Fig. 4. Interface do jogo de decisão sequencial.

(a) Mapa dos edifícios workshop com ações acessíveis ao clicar. (b) Workshop B, onde tampo de mesa e encostos de cadeira são construídos com recursos indicados.

**Idioma:** O estudo foi realizado em alemão

**Construção em partes:** Os participantes construíam partes de móveis, não o item inteiro de uma só vez

**Caminhos sequenciais:** As decisões eram tomadas em sequência, sem aleatoriedade

**Importância do planeamento:** Como os resultados só eram visíveis no final, o planeamento antecipado era fundamental

**Objetivo:** Maximizar os lucros com os recursos limitados disponíveis

**Estrutura do jogo:** Havia 11 tentativas e cada tentativa tinha um limite de 3 minutos

# PRINCIPAIS RESULTADOS

- 01
- 02
- 03
- 04**
- 05
- 06
- 07

## Desempenho Geral:

- A média do desempenho ótimo foi de -10,78 (s.d.: 6,54).
- As soluções foram consideradas completas quando não havia mais móveis possíveis de serem construídos, e incompletas caso contrário.
- Em média, 71,84% das soluções foram completas (s.d.: 9,02).

## Comparação com Agentes Aleatórios:

- Os participantes tiveram desempenho melhor que um agente aleatório ( $p < 0,001$ ) em todas as tentativas, exceto na de treino.

## Estratégias Utilizadas:

- As principais estratégias foram o equilíbrio de recursos, lucro e custo-benefício.

## Correlações:

- Estratégias relacionadas com os recursos disponíveis (tempo e materiais) ocorreram com maior frequência.

## Similaridade com Estratégias Formais:

- A similaridade média entre as escolhas dos participantes e as estratégias formais foi de 0,62 (s.d.: 0,23).
- Estratégias de lucro imediato e custo-benefício corresponderam a 37-55% das escolhas.

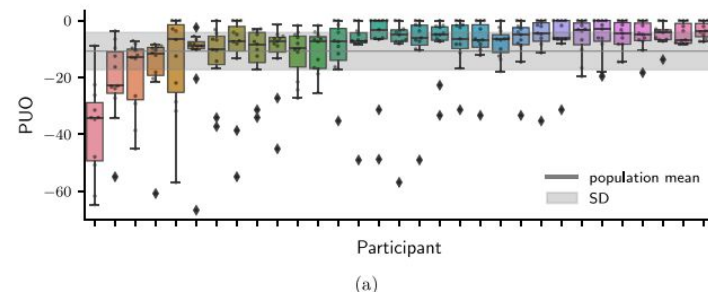


Fig. 5. Desempenho dos participantes na tarefa de decisão sequencial sem backtracking. (a) Percentagem de pontuações abaixo do ótimo, agrupadas por participantes, com média populacional (linha cinza) e desvio padrão (sombreamento).

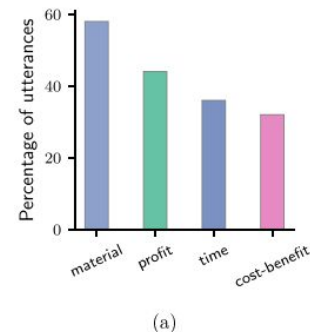


Fig. 6. Medidas de complexidade para as estratégias nas explicações pós-hoc. (a) Percentagem de diferentes estratégias mencionadas.



## PRINCIPAIS CONCLUSÕES DO ESTUDO 2

- 01
  - 02
  - 03
  - 04**
  - 05
  - 06
  - 07
- #1** As estratégias mencionadas pelos participantes foram semelhantes às heurísticas utilizadas no estudo 1.
  - #2** As estratégias formalizadas explicaram uma elevada proporção das escolhas feitas pelos participantes, indicando a sua relevância na modelação do comportamento humano em cenários de otimização.
  - #3** Estas estratégias, baseadas principalmente no equilíbrio de recursos, lucro e custo-benefício, revelaram-se eficazes como modelos gerais para compreender como as pessoas abordam problemas de otimização com restrições.

**ESTUDO 3**  
**(ADAPTAÇÃO DO ESTUDO 2**  
**-> POST-HOC EXPLANATIONS)**

## ESTUDO 3 (ADAPTAÇÃO DE POST-HOC EXPLANATIONS)



Fig. 4. Interface do jogo de decisão sequencial.

(a) Mapa dos edifícios workshop com ações acessíveis ao clicar. (b) Workshop B, onde tampos de mesa e encostos de cadeira são construídos com recursos indicados.

**Objetivo:** Testar como representações e heurísticas se combinam com dados comportamentais.

**Método:** Jogo adaptado do estudo anterior, com algumas mudanças.

### Configurações principais:

- Mudança do jogo para inglês de forma a ter maior alcance
- 6 problemas por participante para reduzir o tempo total
- Variação na ordem das instâncias
- 4 grupos equilibrados, com 12 instâncias distribuídas entre eles, garantindo diversidade nas soluções ótimas
- Instâncias que podem ser resolvidas com estratégias de lucro imediato ou custo-benefício
- Mudança na construção de móveis para simplificar o processo
- Feedback visual com medalhas para indicar o desempenho
- Ações do jogo para análise posterior

# PRINCIPAIS RESULTADOS

## Desempenho:

- Média PUO: -18.72, reduzida para -8.51 considerando apenas soluções completas
- 52% das soluções foram completas

## Semelhança com Estratégias:

- Similaridade média: 0.87
- Melhor correspondência: Instância 2 (0.96)
- Pior correspondência: Instâncias 3 (0.74) e 4 (0.77)
- Estratégias formalizadas descrevem bem as decisões, mas não refletem todo o processo mental

## Análise de Decisões:

- Acordo maior nas decisões dos primeiros passos do jogo
- 57.4% das decisões foram correspondidas por uma ou duas estratégias; 18.7% não corresponderam a nenhuma
- Estratégias mais usadas: Balance material, material gap reduction e immediate profit, variando ao longo do jogo

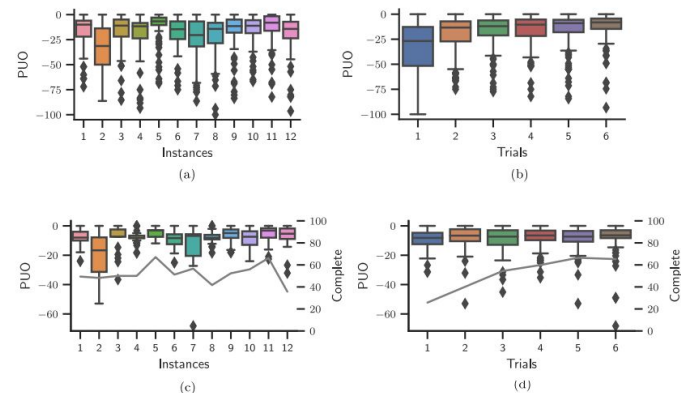


Fig. 7. Percentagens abaixo do ótimo para o conjunto de validação. (a) Por instâncias. (b) Por ensaios. (c) Por instâncias, apenas soluções completas. (d) Por ensaios, apenas soluções completas. As linhas cinzas em (c) e (d) indicam a percentagem de soluções completas.

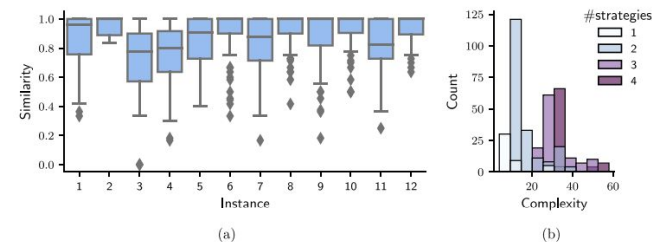


Fig. 8. Similaridade e complexidade das estratégias associadas às trajetórias dos participantes. (a) Similaridade das estratégias por instância. (b) Frequência da complexidade das combinações de estratégias com similaridade 1.



## PRINCIPAIS CONCLUSÕES DO ESTUDO 3

- 01
  - 02
  - 03
  - 04
  - 05**
  - 06
  - 07
- #1** A melhoria do design e uma amostra mais diversificada permitiram análises quantitativas mais detalhadas
- #2** As estratégias formalizadas nos estudos anteriores (1 e 2) foram validadas, com boa correspondência às decisões observadas
- #3** Estratégias principais identificadas:
- Priorização do Lucro
  - Equilíbrio de Recursos
  - Redução do Gap de Materiais

A prevalência dessas estratégias variou ao longo do processo de decisão, destacando padrões dinâmicos

## **WRAP-UP**

## WRAP-UP

- 01
- 02
- 03
- 04
- 05
- 06**
- 07

#1

As estratégias heurísticas humanas, como lucro imediato, balanço, redução de gaps e custo-benefício, podem ser formalizadas para criar explicações cognitivamente alinhadas aos problemas de optimização.



## WRAP-UP

- 01
- 02
- 03
- 04
- 05
- 06**
- 07

#2

As explicações humanas centram-se em subconjuntos de informação relevantes e combinando várias estratégias, sugerindo que os sistemas XAI devem adaptar-se a estratégias humanas para criar narrativas lógicas.



## WRAP-UP

- 01
- 02
- 03
- 04
- 05
- 06**
- 07

#3

As explicações devem ser adaptadas ao público-alvo. Estratégias baseadas no comportamento humano são úteis para utilizadores gerais, enquanto explicações mais técnicas são recomendadas para especialistas.

## WRAP-UP

- 01
- 02
- 03
- 04
- 05
- 06**
- 07

#4

As estratégias humanas formalizadas são independentes do 'solver' e permitem explicações post-hoc.

# LIMITAÇÕES E TRABALHOS FUTUROS

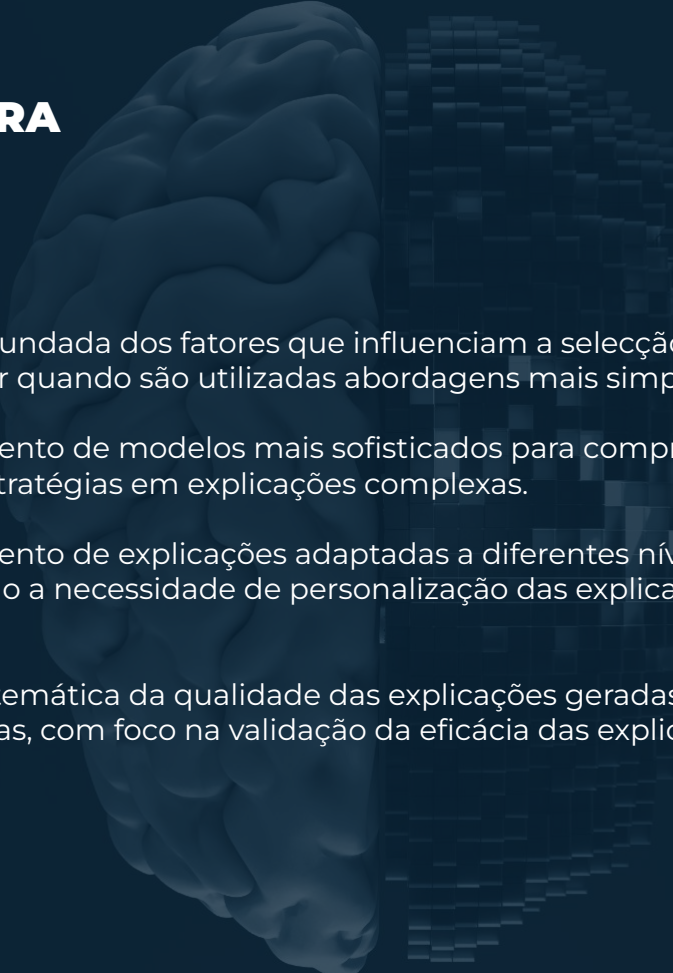
# LIMITAÇÕES

- 01
- 02
- 03
- 04
- 05
- 06
- 07

- #1** A identificação dos gatilhos que determinam a escolha de estratégias específicas revelou-se problemática. Os investigadores depararam-se com dificuldades em determinar quais as características exactas que levam à selecção de uma estratégia em particular.
- #2** Os resultados demonstraram uma correspondência apenas parcial entre as explicações teóricas e as decisões observadas. Em diversos casos, esta correspondência ficou aquém do esperado, sugerindo possíveis lacunas no modelo teórico proposto.
- #3** O estudo demonstrou limitações na modelação completa do raciocínio dos participantes. Os dados recolhidos não foram suficientemente abrangentes para permitir uma compreensão total do processo de raciocínio utilizado.



# INVESTIGAÇÃO FUTURA

- 
- 01
  - 02
  - 03
  - 04
  - 05
  - 06
  - 07**
- #1** Análise aprofundada dos fatores que influenciam a selecção de estratégias, procurando compreender quando são utilizadas abordagens mais simples ou mais complexas.
  - #2** Desenvolvimento de modelos mais sofisticados para compreender a combinação de diferentes estratégias em explicações complexas.
  - #3** Desenvolvimento de explicações adaptadas a diferentes níveis de conhecimento, reconhecendo a necessidade de personalização das explicações consoante o público-alvo.
  - #4** Avaliação sistemática da qualidade das explicações geradas por ferramentas automatizadas, com foco na validação da eficácia das explicações produzidas.



## REFERÊNCIAS BIBLIOGRÁFICAS

- 01 [1] Bassan, S., & Katz, G. (2023). *Towards formal XAI: Formally approximate minimal explanations of neural networks*. In Tools and Algorithms for the Construction and Analysis of Systems (TACAS 2023) (pp. 187–207). Springer.
- 02 [2] Ibs, I., Ott, C., Jäkel, F., & Rothkopf, C. A. (2024). *From human explanations to explainable AI: Insights from constrained optimization*. Cognitive Systems Research, 88
- 03 [3] Waddah Saeed, Christian Omlin. *Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities*. Knowledge-Based Systems. Volume 263. 2023. ISSN 0950-7051
- 04 [4] Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). *Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions*. Information Fusion, 106, 102301
- 05 [5] Cambria, E., Malandri, L., Mercurio, F., Mezzanzanica, M., & Nobani, N. (2023). *A survey on XAI and natural language explanations*. Information Processing & Management, 60(1)
- 06 [6] Chiaburu, T., Haußer, F., & Bießmann, F. (2024). *Uncertainty in XAI: Human perception and modeling approaches*. Machine Learning and Knowledge Extraction, 6(2), 1170–1192.
- 07

**OBRIGADO!**