Licenciatura em Ciência de Dados - 3º ano

From human explanations to explainable AI: Insights from constrained optimization

Inteligência Artificial Simbólica

31 de dezembro de 2024

Discentes: João Dias nº 110305 / David Franco nº 110733

2024/2025

isctе INSTITUTO UNIVERSITÁRIO DE LISBOA

Índice

1. Introdução à Inteligência Artificial Explicativa

1.1 O que é XAI?

Nos últimos anos, a Inteligência Artificial (IA) tem alcançado avanços significativos, consolidando-se como uma ferramenta indispensável em diversas áreas, como saúde, finanças e justiça. Os sistemas baseados em aprendizagem profunda destacam-se pelo desempenho excecional em tarefas complexas, como o reconhecimento de padrões e a previsão de eventos. Contudo, a elevada complexidade desses modelos suscita preocupações quanto à sua transparência e confiabilidade, especialmente quando as suas decisões impactam diretamente a vida das pessoas. Este contexto veio reforçar a necessidade urgente de desenvolver abordagens de Inteligência Artificial Explicativa (XAI – Explainable Artificial Intelligence) [1][2].

A XAI refere-se a um conjunto de técnicas e metodologias que procuram tornar os modelos de IA mais transparentes e compreensíveis para os seres humanos. A explicabilidade é fundamental para garantir que as decisões tomadas por este tipo de sistemas são passíveis de ser justificadas. Nos modelos de "caixa preta", como as redes neuronais, associadas à aprendizagem profunda, as decisões são frequentemente difíceis de interpretar, uma vez que os algoritmos subjacentes são complexos e o processo que conduz a uma determinada decisão não é visível [4].

O desenvolvimento de explicações claras e precisas sobre as decisões dos modelos de IA representa um grande desafio. Este problema deve-se à natureza intrinsecamente complexa dos modelos modernos de IA, que frequentemente lidam com grandes volumes de dados e processos altamente não lineares, o que torna difícil determinar a contribuição de cada variável ou fator nas decisões finais [5][6]. Além disso, questões relacionadas com confiabilidade e responsabilidade em IA tornam a explicabilidade não apenas um requisito técnico, mas uma necessidade ética e legal, especialmente quando os sistemas de IA são usados em contextos sensíveis e críticos [7].

A confiança nas decisões da IA é particularmente importante em domínios críticos, onde os erros podem ter repercussões severas. Por exemplo, em saúde, um

modelo de IA que não consiga explicar as suas decisões pode conduzir a diagnósticos errados ou a tratamentos inadequados. Em finanças, a falta de explicação nas decisões de crédito ou de investimentos pode gerar desconfiança entre os utilizadores e afetar negativamente uma determinada carteira de clientes. Além disso, a ausência de explicações pode permitir a presença de enviesamente nos sistemas, o que pode resultar em discriminação e injustiça, afetando negativamente as pessoas que dependem desses sistemas para tomar decisões [8].

Desta forma, a explicabilidade pode associar-se a uma questão de responsabilidade social e ética, exigindo não só a eficácia/precisão dos modelos de IA, mas também a sua compreensibilidade por parte dos utilizadores, para que possam entender os motivos por trás de decisões automatizadas. A transparência e a confiança são, portanto, essenciais para garantir a adoção responsável e segura da IA em áreas críticas, onde as implicações podem ser profundas e duradouras [2][5].

1.2 Principais técnicas em XAI

Após a explicação do conceito de XAI, e ainda que o objetivo deste trabalho esteja relacionado com problemas de otimização com restrições, é importante aprofundar alguns dos métodos utilizados para tornar as decisões dos modelos mais transparentes e compreensíveis. Existem várias abordagens para atingir esse objetivo, onde cada uma apresenta as suas próprias características e vantagens.

Um dos métodos de destaque é o *Fuzzy Classifier*, que se baseia na lógica *fuzzy* para simular a tomada de decisão humana em situações de incerteza. Este método utiliza quatro componentes principais: a base de conhecimento, que armazena regras condicionais definidas por especialistas; o *fuzzifier*, que transforma os dados de entrada em conjuntos *fuzzy*; o *fuzzy inference engine*, que gera os conjuntos de saída fuzzy a partir das regras; e o *de-fuzzifier*, que converte os valores fuzzy em resultados nítidos. O Fuzzy Classifier é considerado um dos métodos mais transparentes em XAI, pelo seu funcionamento simples e explicável [8].

Outro método importante é o *Gradient-weighted Class Activation Mapping* (Grad-CAM), que é amplamente utilizado em modelos de redes neurais convolucionais (CNN). O Grad-CAM permite gerar mapas de calor que destacam as regiões de uma imagem que são mais relevantes para a decisão do modelo, baseando-se nas informações de gradiente da camada de convolução final. É especialmente útil em aplicações de diagnóstico médico, onde a explicabilidade das previsões é fundamental. No entanto, o Grad-CAM possui algumas limitações, como a dificuldade de localizar objetos que aparecem em várias partes de uma imagem e a perda de sinal durante o processo de *upsampling*. Uma variação melhorada, o Grad-CAM++, resolve algumas dessas limitações e melhora a visualização de classificações multi-label, tornando-se mais confiável para modelos mais complexos [8][10].

O *Layer-wise Relevance Propagation* (LRP) é outro método utilizado para gerar mapas de calor, similar ao Grad-CAM, mas com uma abordagem diferente. O LRP trabalha de forma "retroativa", calculando o score de relevância para cada camada da rede até chegar à entrada do modelo. Este processo permite entender quais as partes da entrada/input layer que contribuíram para a previsão final. O LRP não depende de gradientes e utiliza regras de redistribuição, como as regras básicas, epsilon e gamma, para calcular a relevância [8][11].

Além destes métodos, o Local Interpretable Model-agnostic Explanations (LIME) é uma técnica que permite gerar explicações locais para decisões de modelos complexos. O LIME altera os dados de entrada e analisa como essas mudanças afetam as previsões do modelo. Esta abordagem permite compreender como o modelo tomou uma decisão específica para uma instância individual. Contudo, o LIME pode ser instável devido à aleatoriedade no processo de modificação dos dados. Para superar esta limitação, surgiu o *Deterministic Local Interpretable Model-agnostic Explanations* (DLIME), que utiliza clustering hierárquico para garantir maior estabilidade nas explicações [8][12].

Os Partial Dependence Plots (PDP) são utilizados para mostrar a relação entre variáveis de entrada e a previsão do modelo, ajudando a entender como diferentes

combinações de *features* afetam o *output*. Este método é amplamente utilizado em explicações post-hoc, oferecendo uma visão clara da importância das variáveis [8].

O *SHapley Additive Explanations* (SHAP) é uma técnica baseada na teoria dos jogos que calcula a contribuição de cada *feature* para a previsão do modelo, sendo eficiente para gerar explicações tanto locais quanto globais e é considerado um dos métodos mais robustos de XAI. Existem duas variações principais do SHAP: o Kernel SHAP, que integra os valores de Shapley com o LIME para explicações locais, e o Deep SHAP, que é uma versão específica para redes neurais profundas. Além disso, o TreeExplainer foi desenvolvido para melhorar a eficiência dos cálculos em modelos baseados em árvores, como o Random Forest e o Gradient Boosting [8].

Estes métodos de XAI são essenciais para aumentar a transparência e a confiabilidade dos modelos de IA, especialmente em áreas críticas, como a saúde. A explicabilidade ajuda os profissionais a confiar nas decisões automatizadas e a melhorar os modelos continuamente. Além disso, os métodos de XAI permitem melhorar a colaboração entre especialistas em IA e outros profissionais, como médicos ou engenheiros, permitindo que todos compreendam e validem as decisões do modelo. Em situações onde os resultados de um modelo não são totalmente confiáveis, a XAI oferece a transparência necessária para melhorar os sistemas e garantir a sua precisão [8].

1.3 Otimização com restrições em XAI

Dentro do contexto de XAI, a resolução de problemas de otimização com restrições desempenha um papel crucial, especialmente quando se trata de criar modelos de IA que sejam não apenas eficazes, mas também transparentes e explicáveis. A otimização com restrições envolve a procura da melhor solução possível para um problema, respeitando um conjunto de restrições que podem ser de natureza técnica, ética, legal ou de recursos [6].

A programação linear e outras técnicas de otimização matemática são frequentemente utilizadas para resolver problemas complexos, com aplicações

práticas em diversas áreas da IA. Estas técnicas são especialmente úteis quando as respostas do sistema precisam de ser apresentadas de forma clara e concisa, mantendo a precisão e otimizando o desempenho. Por exemplo, ao resolver problemas de gestão de recursos ou alocação de tarefas em sistemas de IA, a programação linear pode ser usada para encontrar a solução ótima dentro das restrições impostas, como limitações de tempo, orçamento ou restrições ambientais. Estes problemas surgem frequentemente em áreas como a gestão de carteiras financeiras ou associados a tarefas de planeamento [5].

O principal desafio da otimização com restrições está relacionado com a complexidade dos Solvers, que embora sejam extremamente poderosos, podem ser bastante complexos e difíceis de interpretar, especialmente quando as soluções envolvem grandes conjuntos de dados, múltiplas restrições e variáveis de decisão. A dificuldade de explicação das soluções torna-se evidente quando os solvers geram resultados em que o raciocínio por trás da escolha da solução não é facilmente acessível aos utilizadores ou não pode ser traduzido em explicações claras para os decisores humanos [7].

2. Estudo de Caso: "From Human Explanations to Explainable AI: Insights from constrained optimization"

Toda esta secção refere-se ao estudo de caso “From Human Explanations to Explainable AI: Insights from constrained optimization” [9]. Para evitar a repetição excessiva de referências, optou-se por omitir a referenciação, uma vez que todas as informações apresentadas nesta secção estão baseadas neste artigo.

O estudo ilustra como a XAI é aplicada no contexto de otimização com restrições e explora três experiências em que são utilizados problemas de otimização com restrições no formato de jogos de computador, todos baseados no paradigma da Furniture Factory, onde os participantes gerem uma empresa de móveis e cujo objetivo é otimizar a produção de móveis, respeitando uma série de restrições, como a capacidade de produção e os custos operacionais.

As soluções ótimas para estes problemas de otimização são obtidas com recurso a Solvers/algoritmos de programação linear que garantem a optimalidade da solução, mas que apresentam desafios quando se trata de explicar essas soluções para “não especialistas”. Embora os algoritmos utilizados sejam bem compreendidos e eficazes na resolução do problema, a complexidade das representações dos problemas num espaço abstrato de alta dimensão torna difícil traduzir as soluções em explicações intuitivas e acessíveis aos utilizadores.

Nos estudos qualitativos realizados, os investigadores analisaram como os participantes formulam explicações para as suas decisões dentro do jogo e como essas explicações podem ser formalizadas em estratégias cognitivamente adequadas. A análise revelou que as explicações fornecidas pelos participantes tendem a variar em complexidade, dependendo de como as soluções ótimas são interpretadas em relação às restrições envolvidas. A complexidade das explicações geradas pelos jogadores foi comparada com a complexidade das estratégias de decisão ótimas e discutiu-se como as explicações cognitivamente adequadas podem ser utilizadas para melhorar a confiança nas soluções de otimização com restrições.

Este estudo de caso demonstra a relevância da explicabilidade nas soluções de otimização com restrições, destacando como a complexidade dos solvers pode tornar as soluções difíceis de entender, mesmo quando a solução é ótima do ponto de vista matemático. A combinação de otimização matemática e técnicas de explicabilidade é fundamental para garantir que as decisões tomadas por sistemas baseados em IA sejam tanto eficientes quanto compreensíveis para os utilizadores.

Além disso, o estudo fornece insights sobre como as heurísticas e estratégias cognitivas utilizadas pelos humanos para resolver problemas complexos de otimização podem ser formalizadas e comparadas com as soluções ótimas geradas por Solvers. Esta abordagem pode ser aplicada para melhorar a interpretação das soluções e ajudar a reduzir a complexidade das explicações, tornando-as mais acessíveis aos utilizadores e aumentando a confiança nas decisões de IA.

2.1 Objetivo do Estudo

O estudo tem como objetivo principal investigar como as explicações humanas podem ser integradas e utilizadas para melhorar a explicabilidade das soluções geradas por sistemas baseados em otimização com restrições. Em particular, ele procura compreender como as estratégias cognitivas que os seres humanos utilizam para resolver problemas de otimização podem ser formalizadas e aproveitadas no desenvolvimento de explicações mais transparentes e interpretáveis em sistemas de inteligência artificial (IA). O estudo propõe que, ao compreender as representações mentais e as heurísticas que guiam as decisões humanas em problemas complexos, podemos criar explicações que não só ajudem os utilizadores a entender como a IA chegou a uma solução, mas também tornem as soluções mais plausíveis e confiáveis, especialmente em cenários de otimização com restrições.

2.2 Metodologia

Para a recolha de dados sobre as explicações dos participantes e as suas estratégias cognitivas, foram utilizados três jogos baseados no paradigma da Furniture Factory, que se encontra detalhado na Fig. 1, onde os participantes eram desafiados a resolver problemas de otimização com restrições no contexto de uma fábrica de móveis. Todos os jogos simulam um ambiente de produção em que os participantes devem tomar decisões relacionadas com o número de recursos que devem ser produzidos de forma a maximizar o lucro e respeitando diversas restrições. Todos os estudos apresentam uma estrutura muito similar, ainda que com algumas nuances importantes entre si, que serão detalhadas nas secções 2.3.1, 2.3.2 e 2.3.3.

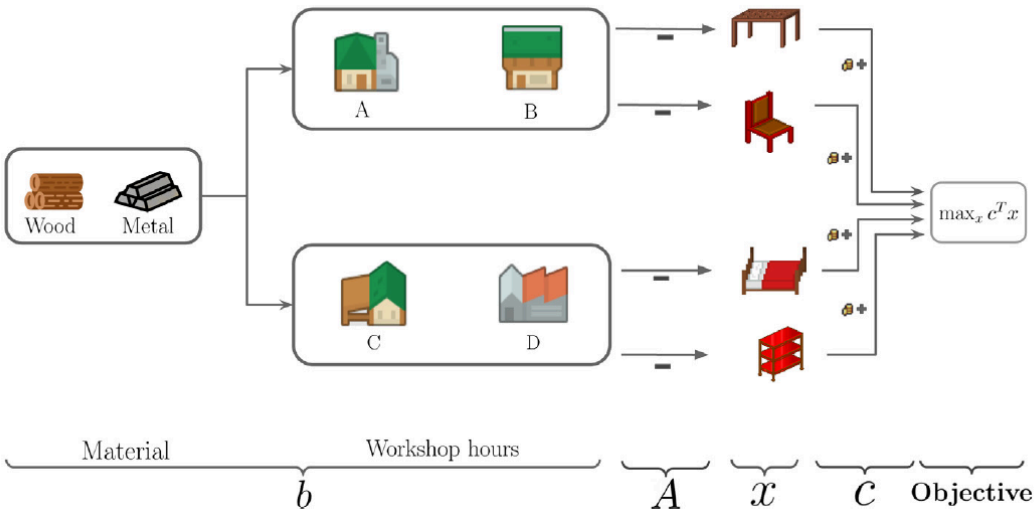

Fig. 1 - Overview do design estrutural do paradigma da Furniture Factory

Este cenário de otimização da Furniture Factory foi escolhido porque, embora envolva alguma complexidade matemática, as decisões dos participantes podem ser guiadas por uma variedade de heurísticas e representações cognitivas que refletem estratégias humanas típicas, como a procura por soluções rápidas ou a análise cuidadosa de custos e benefícios. Durante a experiência, os participantes são incentivados a explicar as estratégias e raciocínios que utilizam para tomar decisões, permitindo que os investigadores capturem essas explicações de forma detalhada.

2.3 Descrição dos Três Estudos Realizados

2.3.1 Estudo 1: Elicitação de Representações e Heurísticas Através de Explicações Simultâneas

O primeiro estudo foi realizado num laboratório universitário, utilizando um jogo de computador, que permitiu observar como os participantes resolvem problemas de otimização com restrições e como explicam as suas decisões ao longo do processo. A tarefa envolvia a gestão da fábrica de móveis, onde os participantes ajustavam controladores deslizantes para determinar a quantidade de camas,

estantes, mesas e cadeiras a produzir, com base em recursos disponíveis, custos de produção e lucros esperados. Os participantes tiveram que gerir a Furniture Factory durante sete tentativas (onde cada uma equivalia a 1 mês) correspondentes a sete problemas distintos, incluindo um tutorial/tentativa inicial. A Fig. 2 demonstra a interface do jogo de computador utilizada, incluindo os elementos interativos para o ajuste das quantidades de móveis construídos (a) e o resumo inicial das alterações no lucro líquido e no material disponível antes de cada tentativa (b).

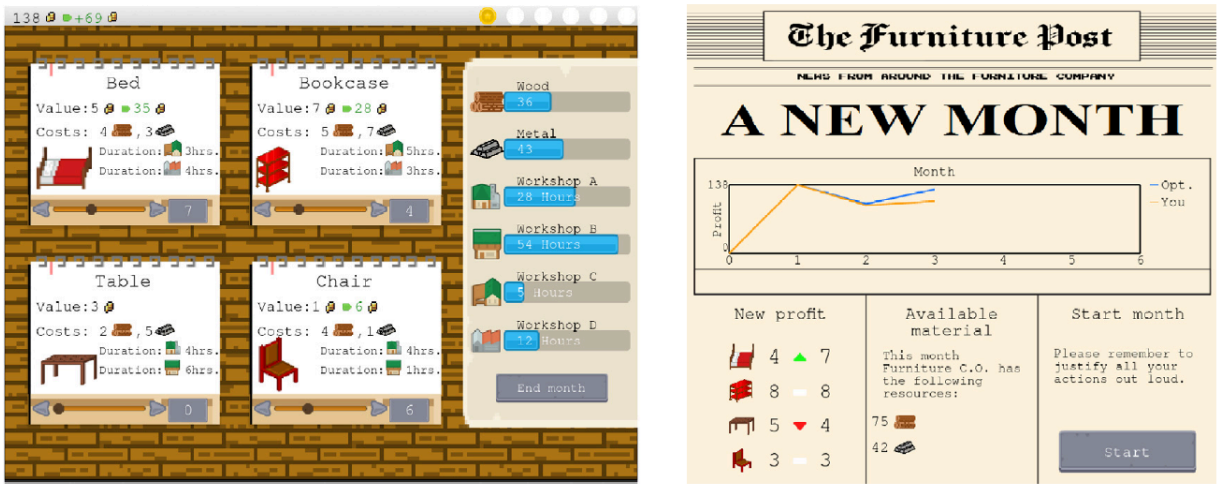

Fig. 2 - Jogo de computador. (a) A interface principal com todas as informações relevantes para a tarefa e sliders para modificar as quantidades de móveis construídos. (b) Ecrã apresentado no início de cada tentativa, indicando as alterações no lucro líquido dos móveis e o material disponível.

As informações essenciais, como recursos, custos e lucros, estavam visíveis no ecrã, e um sistema de estrelas indicava soluções ótimas. Os utlizadores recebiam um feedback comparativo com a solução ótima, após cada tentativa, e foram divididos em dois grupos onde era pedido para justificar verbalmente as suas ações durante o jogo. As interações foram gravadas e transcritas para análise qualitativa. A análise identificou três estratégias principais: equilíbrio de recursos, orientação para o lucro e o custo-benefício. Observou-se que os participantes se focavam em subconjuntos de atributos, evidenciando a simplicidade na representação do problema. Doze utilizadores participaram, no entanto um foi excluído devido a problemas técnicos,

resultando em 11 participantes com idades entre os 18 e 29 e experiência variada em jogos.

Os resultados indicaram que o desempenho médio dos participantes foi próximo do ótimo, -2,0% pelo Percentual Abaixo do Ótimo (PUO). Nenhum participante alcançou a solução ótima e o desempenho variou entre os todos os utilizadores. A análise das explicações mostrou que as estratégias de equilíbrio de recursos e custo-benefício foram as mais comuns em tentativas posteriores, enquanto a maximização de lucros teve mais predominância no início. Este estudo evidenciou que os participantes utilizam abordagens estruturadas para otimização sob restrições, o que pode informar o desenvolvimento de explicações para sistemas de XAI.

2.3.2 Estudo 2: Validação das Explicações Post-Hoc e Análise das Representações Cognitivas

Neste segundo estudo, o objetivo dos participantes era resolver um problema de otimização onde tinham de construir móveis a partir de recursos limitados, sem a possibilidade de reverter as suas decisões, ao contrário do primeiro estudo. Desta forma, pretendia-se entender se as representações do problema e as estratégias usadas pelos participantes se aplicariam a uma tarefa de decisão sequencial. A Fig. 3 demonstra a interface do jogo para a tarefa de tomada de decisão sequencial, incluindo o mapa dos edifícios (a), onde as ações e informações são acedidas através de cliques, e a Fábrica B (b), onde tampos de mesa e encostos de cadeiras são construídos com os recursos indicados, permitindo a organização de itens em fila para construção sequencial.

Fig. 3 - Interface do jogo para a tarefa de tomada de decisão sequencial. (a) Interface principal com o mapa dos diferentes edifícios. As ações e informações são acedidas ao clicar nos diferentes edifícios. (b) Fábrica B, onde podem ser construídos tampos de mesa e encostos de cadeiras com os recursos indicados

A performance média dos participantes foi inferior à solução ótima e registaram-se 71,84% das soluções completas. Ao comparar com um agente aleatório, os participantes obtiveram resultados significativamente melhores em todas as tentativas, à exceção do treino. As explicações dadas pelos participantes foram analisadas e agrupadas em várias estratégias, como equilibrar os recursos, escolher com base no lucro e avaliar a relação custo-benefício. Essas estratégias refletiram um raciocínio que combinava decisões simples com decisões mais complexas, envolvendo tanto os recursos disponíveis quanto os custos associados. Essas estratégias foram formalizadas em quatro categorias principais: lucro imediato, equilíbrio de recursos, redução do gap de materiais e custo-benefício. Cada uma reflete como os participantes poderiam ter agido caso tivessem recursos computacionais para realizar cálculos exatos. Estas formalizações ajudam a descrever o raciocínio dos participantes, embora estes não realizassem cálculos precisos. Por exemplo, a estratégia de "lucro imediato" escolhe o item com maior lucro, enquanto o "equilíbrio" tenta equilibrar os recursos disponíveis, quer seja em termos de materiais, quer seja em termos de tempo.

O estudo confirma que as estratégias usadas pelos participantes podem ser formalizadas para descrever eficazmente as suas decisões. Apesar de haver variações nas escolhas, essas estratégias conseguem captar bem o raciocínio dos participantes, tornando-se uma base útil para descrever como os problemas de otimização são resolvidos.

2.3.3 Estudo 3: Análise de Dados Comportamentais e Comparação das Estratégias Cognitivas com a Complexidade das Explicações Fornecidas

No terceiro estudo, procurou-se explorar como as estratégias formalizadas se alinham com as escolhas dos participantes, utilizando uma plataforma online com uma amostra maior e diversificada. O objetivo foi validar as estratégias identificadas nos estudos anteriores e obter mais detalhes quantitativos sobre o uso dessas estratégias. Nesta etapa, o jogo pedia a resolução de seis instâncias do problema, com tempo reduzido e algumas adaptações no jogo, como a introdução de medalhas para premiar os participantes com melhores soluções.

A amostra final teve 167 participantes com idades entre os 18 e os 60, onde a maioria jogava regularmente. É importante notar que foram excluídos 40 dos 207 participantes recrutados por não apresentarem soluções completas. Os resultados mostraram que a média da pontuação de desempenho ficou a 18,72 pontos do valor ótimo (PUO), com 52% das soluções completas (inferior ao valor apresentado no segundo estudo). Ao considerar apenas as soluções completas, a pontuação média foi de -8,51. A análise de similaridade das estratégias mostrou que as estratégias formalizadas correspondiam bem às decisões dos participantes, com uma pontuação média de 0,87. As estratégias mais eficazes para descrever as decisões dos participantes traduziram-se no equilíbrio de materiais, na redução do gap de materiais e no lucro imediato.

Por fim, a análise dos pontos de decisão revelou que os participantes tendiam a concordar mais nas escolhas das etapas iniciais das trajetórias. Quando várias estratégias coincidiam nas mesmas escolhas, isso apontava logo que diferentes

abordagens podiam levar à mesma decisão, tornando as estratégias acima mencionadas ainda mais importantes.

3. Implicações para XAI

Os resultados obtidos têm implicações significativas para o campo da XAI, especialmente em sistemas de otimização com restrições. Os resultados sugerem que as explicações geradas por IA devem adotar estratégias cognitivas semelhantes às usadas por humanos, refletindo as heurísticas aplicadas na resolução de problemas complexos. Sublinha-se também a necessidade de ajustar os métodos de explicação de IA, garantindo que sejam não só precisos, mas também adaptados ao público-alvo. Este alinhamento tornará as explicações mais intuitivas e compreensíveis, pois partilham a mesma base cognitiva.

No estudo é demonstrado que as explicações humanas para problemas de otimização com restrições utilizam com frequência uma combinação de estratégias cognitivas, como a procura pelo lucro mais imediato e a análise de custo-benefício. Essas estratégias podem ser formalizadas e utilizadas como blocos de construção para gerar explicações automatizadas, tornando-as mais alinhadas ao processo de tomada de decisão humano. Esta abordagem não só melhora a compreensão dos resultados das IA, mas também garante que as explicações sejam relevantes e úteis, independentemente da complexidade do problema em questão.

Essa adaptação das explicações geradas por IA, com base em métodos humanos, não só torna as soluções mais acessíveis, mas também promove a confiança no sistema, uma vez que os utilizadores podem entender as decisões tomadas pela IA de uma maneira mais próxima à forma como eles próprios raciocinam.

4. Conclusões

O estudo que foi abordado ao longo deste relatório revela a importância de integrar as explicações humanas no desenvolvimento de sistemas de XAI, especialmente no contexto de otimização com restrições. Ao analisar como os participantes resolvem problemas de otimização e explicam as suas escolhas, observou-se que as estratégias cognitivas utilizadas podem ser formalizadas em prol de refletir com precisão o raciocínio presente nas decisões. Estas estratégias demonstraram ser eficazes na descrição de como os humanos abordam problemas complexos, mesmo sem recorrer a cálculos exatos. Ao juntar as explicações de IA com essas abordagens cognitivas, é possível oferecer soluções mais intuitivas e também mais confiáveis para os utilizadores. Assim, é essencial incorporar estas estratégias nos sistemas de IA para melhorar a transparência, facilitar a compreensão das soluções e aumentar a confiança dos utilizadores nas decisões tomadas por esses sistemas.

Referências

[1] Bassan, S., Katz, G. (2023). Towards Formal XAI: Formally Approximate Minimal Explanations of Neural Networks. In: Sankaranarayanan, S., Sharygina, N. (eds) Tools and Algorithms for the Construction and Analysis of Systems. TACAS 2023. Lecture Notes in Computer Science, vol 13993. Springer, Cham. https://doi.org/10.1007/978-3-031-30823-9_10

[2] Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. Knowledge-Based Systems, 263, 110273. https://doi.org/10.1016/j.knosys.2023.110273

[3] 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., & others. (2024). Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary

research directions. Information Fusion, 106, 102301. https://doi.org/10.1016/j.inffus.2024.102301

[4] Chiaburu, T., Haußer, F., & Bießmann, F. (2024). Uncertainty in XAI: Human perception and modeling approaches. Machine Learning and Knowledge Extraction, 6(2), 1170–1192. https://doi.org/10.3390/make6020055

[5] Cambria, E., Malandri, L., Mercorio, F., Mezzanzanica, M., & Nobani, N. (2023). A survey on XAI and natural language explanations. Information Processing & Management, 60(1), 103111. https://doi.org/10.1016/j.ipm.2022.103111

[6] State, L., Ruggieri, S., Turini, F. (2023). Declarative Reasoning on Explanations Using Constraint Logic Programming. In: Gaggl, S., Martinez, M.V., Ortiz, M. (eds) Logics in Artificial Intelligence. JELIA 2023. Lecture Notes in Computer Science(), vol 14281. Springer, Cham. https://doi.org/10.1007/978-3-031-43619-2_10

[7] Maddigan, P., Lensen, A., & Xue, B. (2024). Explaining genetic programming trees using large language models. arXiv:2403.03397. https://arxiv.org/abs/2403.03397

[8] Saranya, A., & Subhashini, R. (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. Decision Analytics Journal, 7, 100230. https://doi.org/10.1016/j.dajour.2023.100230

[9] Ibs, I., Ott, C., Jäkel, F., & Rothkopf, C. A. (2023). From human explanations to explainable AI: Insights from constrained optimization. *Cognitive Science & Institute of Psychology*, TU Darmstadt.

[10] Linardatos P., Papastefanopoulos V., Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. Entropy, 23 (2021), p. 18, https://doi.org/10.3390/e23010018

[11] Loh Hui Wen, Ooi Chui Ping, Seoni Silvia, Barua Prabal Datta, Molinari Filippo, Rajendra Acharya U. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). Comput. Methods Programs Biomed., 226, https://doi.org/10.1016/j.cmpb.2022.107161

[12] Zafar Muhammad Rehman, Khan Naimul Mefraz (2019). DLIME: A deterministic local interpretable model-agnostic explanations approach for computer-aided diagnosis systems. arXiv preprint arXiv:1906.10263. https://arxiv.org/abs/1906.10263