REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

A EVOLUÇÃO DA INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (IAE)

THE EVOLUTION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

LA EVOLUCIÓN DE LA INTELIGENCIA ARTIFICIAL EXPLICABLE (IAE)

Juliano Araujo Santana[1], Angelo Machado de Souza[2], Michel Souza Silva[3], Davis Souza Alves[4], Márcio Magera Conceição[5]

e737438
https://doi.org/10.47820/recima21.v7i3.7438

PUBLICADO: 03/2026

RESUMO

Este artigo analisa a evolução da Inteligência Artificial (IA) até a consolidação da Inteligência Artificial Explicável (XAI), enfatizando como o avanço do *machine learning* e, sobretudo, do *deep learning*, aumentou o desempenho dos modelos ao custo de maior opacidade ("caixa-preta"). O estudo evidencia a aplicabilidade da IA nas áreas de saúde, finanças, justiça, segurança e outros contextos críticos, nos quais decisões automatizadas podem afetar direitos fundamentais. Nesses domínios, torna-se imprescindível que os sistemas sejam confiáveis, auditáveis e responsáveis, reforçando a necessidade de explicações claras e justificáveis. A revisão da literatura destaca a importância das explicações à medida que a XAI evolui de abordagens *post hoc*, como LIME e SHAP, para a sistematização teórica dos conceitos de interpretabilidade e explicabilidade. Quanto

[1] Mestrando em Administração com pesquisa em Inteligência Artificial Explicável (XAI), pós-graduado em Gestão de TI e Cloud Computing (UFSCar) e graduado em Gestão de TI (UNIP), com certificações internacionais em ITIL® 4, Security+, DPO e ISO 27001. Analista de Suporte Pleno na Stefanini Brasil (Bat Latam South) e como Consultor de TI Jr., É avaliador acadêmico em cursos de TI e Coordenador Nacional do Comitê Startech da APDADOS.

[2] Mestrado em Administração pela Florida Christian University com pesquisa em Novas Tecnologias Avançadas. MBA em Gerenciamento de Projetos de TI e certificações internacionais, como Certified Information Security Officer (CISO), ISO/IEC 27001 Lead Implementer e DPO – EXIN. Graduado em Tecnologia de Redes de Computadores. Experiência consolidada na liderança de projetos em empresas como NTT Ltd., ISH Tecnologia e Grupo Cornélio Brennand.

[3] Tecnólogo em Marketing (UNIP) e pós-graduado em Data Protection Officer (LGPD/GDPR), com especialização em Social Media e atuação na gestão de redes sociais para profissionais de TI e instituições do terceiro setor. É Gerente de Marketing e Assessor do Comitê Diretivo da APDADOS, com atuação institucional em Brasília junto a órgãos federais entre 2021 e 2023. No cenário internacional, realizou missões oficiais em países como França, Inglaterra e Angola, onde foi palestrante convidado pelo Ministro de Tecnologia.

[4] Doutor em Administração de TI - Ph.D pela Florida Christian University (EUA) convalidado no Brasil, Mestre em Administração com foco em TI Verde (2015), Extensão em Gestão de TI pela FGV/SP (2011), Pós-graduado em Gerenciamento de Projetos (2009). Professor de Segurança da Informação na Universidade Paulista (UNIP), Universidade Municipal de São Caetano do Sul (USCS) e Flórida Christian University (FCU). Possui as certificações PMP®, ITIL® Expert, C|EH®, C|HFI® e EXINI® Data Protection Atua nos Estados Unidos como Gerente de Projetos de Cibersegurança (R&D) com foco em Privacidade de Dados (LGPD/GDPR), Computação Forense, Ethical Hacker e Inteligência Artificial(AI). Presidente da Associação Nacional dos Profissionais de Privacidade de Dados (APDADOS.org).

[5] Doutor em Economista pela PUC - Campinas. MBA de Marketing – ESAMC, Sorocaba. Mestrado em Administração pela UNG -Guarulhos. Mestrado em Sociologia pela PUC -São Paulo. Doutorado em Sociologia pela PUC -São Paulo. Doutorado em Administração pela FCU -USA. Pós Doutorado Unicamp -Campinas. Pós Doutor FCU -USA. Pós Doutor UC-Portugal. Jornalista e Escritor. Avaliador do MEC/INEP. Pró Reitor da Universidade Guarulhos, SP.

ao método, o estudo combina revisão narrativa, exame bibliográfico e documental (incluindo análises regulatórias relacionadas ao GDPR) e investigação bibliométrica. Os resultados bibliométricos (Scopus, 2004–2023) indicam crescimento expressivo das publicações sobre XAI a partir de 2018, identificando autores centrais e principais fontes editoriais, como revistas e conferências. Entretanto, considerando a natureza ambígua do termo "XAI" nos critérios de busca, observa-se que os dados do Google Trends confirmam o aumento do interesse público por "inteligência artificial explicável", o que exige cautela metodológica na condução das consultas. Por fim, a XAI configura-se como uma resposta técnico-sociotécnica à complexidade dos modelos de IA, às pressões éticas e às crescentes exigências regulatórias por transparência.

PALAVRAS-CHAVE: Inteligência Artificial Explicável. XAI. Transparência Algorítmica. Auditoria Algorítmica. Viés Algorítmico.

ABSTRACT

This article analyzes the evolution of Artificial Intelligence (AI) up to the consolidation of Explainable Artificial Intelligence (XAI), emphasizing how the advancement of machine learning and, above all, deep learning has increased the performance of models at the cost of greater opacity ("black box"). The literature review shows that explainability became essential as AI systems were adopted in high-stakes domains (healthcare, finance, justice, and security), where automated decisions affect rights and require trust, auditability, and accountability. Key XAI milestones include post-hoc explanation methods such as LIME and SHAP and conceptual frameworks that distinguish interpretability and explainability. Methodologically, the study combines a narrative literature review, bibliographic and documentary analysis (including regulatory discussions related to the GDPR), and bibliometric analysis. Results based on bibliometric evidence (Scopus, 2004–2023) indicate rapid growth of XAI publications from 2018 onward, identifying leading authors and major publication venues (journals and proceedings series). In addition, Google Trends suggests rising public interest in "explainable artificial intelligence," while also revealing semantic ambiguity around the term "XAI" in web searches, which can bias query-based analyses unless carefully controlled. Overall, XAI emerges as a technical and socio-technical response to model complexity and to ethical and regulatory demands for transparency.

KEYWORDS: *Explainable Artificial Intelligence. XAI. Algorithmic Transparency. Algorithmic Auditing. Algorithmic Bias.*

RESUMEN

Este artículo analiza la evolución de la Inteligencia Artificial (IA) hasta la consolidación de la Inteligencia Artificial Explicable (XAI), enfatizando cómo el avance del machine learning y, sobre todo, del deep learning incrementó el rendimiento de los modelos a costa de una mayor opacidad ("caja negra"). El estudio evidencia su aplicabilidad en los ámbitos de la salud, las finanzas, la justicia, la seguridad y otros contextos críticos en los que las decisiones automatizadas afectan derechos; por lo tanto, la IA se utiliza en estos dominios y, en consecuencia, debe ser confiable, auditable y responsable, lo que subraya la necesidad de explicaciones en dichos entornos. Una revisión de la literatura pertinente destaca la importancia de las explicaciones a medida que la XAI evoluciona desde enfoques post-hoc, como LIME y SHAP, al tiempo que el concepto de interpretabilidad/explicabilidad se sistematiza progresivamente en el plano teórico. En cuanto a la metodología empleada, este estudio combina revisión narrativa, análisis bibliográfico y documental (incluyendo discusiones regulatorias sobre el GDPR) y análisis bibliométrico. Las evidencias bibliométricas (Scopus, 2004–2023) indican un crecimiento exponencial de publicaciones sobre XAI a partir de 2018, identificándose los autores centrales y las principales fuentes editoriales (revistas y series de conferencias). No obstante, debido al carácter ambiguo del término "XAI" en las búsquedas, se observa que los datos de Google Trends confirman un aumento del interés público por la "inteligencia artificial explicable", lo que exige cautela metodológica en la interpretación de las consultas. Finalmente, la XAI representa una respuesta técnico-sociotécnica

a la creciente complejidad de los modelos, a las presiones éticas y a las demandas regulatorias cada vez más intensas en favor de la transparencia.

PALABRAS CLAVE: *Inteligencia artificial explicable. XAI. Transparencia algorítmica. Auditoría algorítmica. Sesgo algorítmico.*

1. INTRODUÇÃO

A inteligência artificial (IA) percorreu um longo caminho desde que foi desenvolvida como um campo científico na década de 1950. Originou-se de abordagens simbólicas e sistemas baseados em regras, mas foi a "IA tradicional", com sua representação explícita de conhecimento do campo, que tornou seus processos de tomada de decisão relativamente transparentes. Modelos estruturados como sistemas especialistas e árvores de decisão permitem uma exploração estruturada de regras e inferências (Russell; Norvig, 2016). No entanto, juntamente com a melhoria progressiva dos métodos estatísticos e do aprendizado de máquina, mais notavelmente desde os anos 2000, o campo testemunhou uma mudança de paradigma de modelos simbólicos para modelos orientados por dados - levando à ascensão do aprendizado de máquina e do aprendizado profundo.

Houve três razões principais para o crescimento do aprendizado profundo: aumento do poder computacional, grande disponibilidade de dados e melhorias nos algoritmos de aprendizado de máquina em redes neurais artificiais (Lecun; Bengio; Hinton, 2015). Por exemplo, essas peças levaram a uma grande melhoria nas tarefas de reconhecimento de imagem, processamento de linguagem natural, sistemas de recomendação. No entanto, a maior complexidade estrutural desses modelos foi acompanhada por uma fraqueza principal: a caixa preta, ou opacidade, desses sistemas (Goodfellow; Bengio; Courville, 2016).

Enquanto nos sistemas simbólicos convencionais era possível ver onde uma conclusão havia sido alcançada a partir da sequência lógica de eventos, modelos que usam redes neurais profundas têm milhões ou até bilhões de parâmetros que são ajustados automaticamente a partir de dados. Essa característica leva a um desafio para o ser humano ao interpretar os resultados de sua tomada de decisão. Como Lipton (2016) argumenta, o significado de "interpretabilidade" ganhou ampla popularidade, mas também tem sido comumente interpretado de forma ambígua, incluindo: transparência estrutural, explicação pós-hoc, capacidade de simulação cognitiva.

O uso crescente de IA em circunstâncias urgentes tornou a explicabilidade ainda mais necessária. Sistemas automatizados começaram a aparecer em diagnósticos médicos, concessão de crédito, contratação, questões judiciais e segurança pública. Nesses contextos, decisões automatizadas podem impactar os direitos das pessoas ou até mesmo de comunidades, e, portanto, é fundamental compreender como e por que determinada decisão foi alcançada (Doshi-

Velez; Kim, 2017). A falta de explicações pode minar a confiança, a responsabilidade e a legitimidade social em relação a essas tecnologias.

Além disso, a literatura indica a capacidade dos modelos de IA de emular ou amplificar preconceitos embutidos no conjunto de dados de treinamento. Barocas e Selbst (2016) mostram que sistemas algorítmicos podem ter efeitos discriminatórios sem a necessidade de criar isso intencionalmente e podem, de fato, reforçar a desigualdade estrutural. Este caso reforçou a demanda por um mecanismo de auditoria e interpretações de decisões automatizadas, por maior abertura e pela justiça de nossos algoritmos.

Nesse sentido, a Inteligência Artificial Explicável *(XAI)* torna-se evidente e se estabelece como um subcampo para desenvolver e construir técnicas para entender modelos complexos. O termo ganhou aceitação mundial a partir de 2016, particularmente através do programa *XAI* da *DARPA* (Gunning, 2017), e seu foco em criar sistemas que explicariam informações para operadores humanos.

Alguns avanços científicos têm facilitado a consolidação da IA explicável. Para esse fim, Ribeiro *et al.,* (2016) apresentaram o método *LIME Local Interpretable Model-Agnostic Explanations* (LIME), que permite que as explicações de decisões individuais sejam dadas em estimativas locais razoáveis pelo respectivo classificador. Lundberg e Lee (2017) então introduziram o SHAP *(SHapley Additive exPlanations)* para desenvolver uma estrutura adaptativa que pode ser usada para considerar atribuições únicas da importância das variáveis em modelos complexos, com base na teoria dos jogos cooperativos.

Além disso, como sugerido por Molnar (2022), os autores fizeram uma tipologia lógica dos métodos de interpretabilidade entre explicações globais e locais, intrínsecas e pós-hoc, fornecendo assim uma base conceitual para a pesquisa. Arrieta *et al.* (2020) oferecem uma extensa revisão da literatura sobre o tema da *XAI*, que examina conceitos incluindo interpretabilidade, explicabilidade, transparência, compreensibilidade e os amplos aspectos interdisciplinares do campo.

O desenvolvimento da IA explicável também foi informado por debates políticos. O Regulamento Geral sobre a Proteção de Dados (GDPR) da União Europeia, emitido em 2018, estimulou discussões sobre o chamado “direito à explicação” em relação a decisões automatizadas e cabe aos Encarregados de Proteção de Dados – *Data Protection Officers* (DPO) incentivar essas discussões nas empresas (Alves, 2021). Embora existam interpretações legais controversas desse direito, Wachter, Mittelstadt e Floridi (2017) consideram a legalidade da transparência algorítmica e a necessidade de mecanismos explicativos para entender a complexidade dos sistemas.

Assim, a mudança da IA tradicional para a IA explicável pode ser interpretada como um composto de quatro eixos interseccionais — (i) sofisticação técnica dos modelos, impulsionada

pelo avanço do *deep learning* (Lecun; Bengio; Hinton, 2015; Goodfellow; Bengio; Courville, 2016), (ii) crescente utilização em contextos sensíveis, que demandam maior interpretabilidade (Doshi-Velez; Kim, 2017), (iii) considerações éticas relacionadas a preconceitos e discriminação algorítmica (Barocas; Selbst, 2016), e (iv) desafios regulatórios em torno da transparência e responsabilidade em decisões automatizadas (Wachter; Mittelstadt; Floridi, 2017). Alguns países, por exemplo, estão à frente nos avanços científicos tanto em IA quanto em XAI, conforme evidenciado por análises bibliométricas recentes que apontam liderança de Estados Unidos, Alemanha, Reino Unido e China na produção científica da área (Sharma *et al.,* 2024; Zhou *et al.,* 2019).

Com base em dados bibliométricos, os Estados Unidos lideram o mundo em número de publicações sobre inteligência artificial, com universidades como MIT, Stanford e Carnegie Mellon liderando esse crescimento. A IA tem visto uma produção científica notavelmente rápida na China ao longo das décadas e agora está tornando-se um centro global chave (Zhou *et al.,* 2019).

Na Europa, países do Reino Unido, Alemanha e França estão na vanguarda da pesquisa que harmoniza inteligência artificial, ética e regulação tecnológica, especialmente no contexto de diretrizes para uma IA confiável e centrada no ser humano (COMISSÃO EUROPEIA, 2019; Sharma *et al.,* 2024). Por exemplo, o Reino Unido é reconhecido por articular o diálogo entre academia e política pública no campo da governança algorítmica, como demonstram as orientações conjuntas do *Information Commissioner's Office (ICO)* e do *Alan Turing Institute* sobre explicação de decisões automatizadas (ICO, 2020; GOV.UK, 2018).

Assim, o estabelecimento da IA explicável representa não apenas um avanço técnico, mas também uma resposta às exigências sociais e institucionais por transparência, confiabilidade e responsabilidade em sistemas automatizados (Wachter; Mittelstadt; Floridi, 2017; COMISSÃO EUROPEIA, 2019). Trata-se de uma mudança paradigmática que desloca o foco exclusivo no desempenho preditivo para incorporar compreensibilidade, prestação de contas e governança como demandas centrais do desenvolvimento e aplicação de modelos de IA.

A questão-problema foi desenvolvida sob essas condições: Quais fatores contribuíram para a evolução da Inteligência Artificial Convencional para a Inteligência Artificial Explicável, bem como os países de destaque?

Em resposta a essa pergunta, o presente artigo visa estudar como a IA explicável se desenvolveu ao longo dos anos, focando nos principais marcos teóricos e metodológicos da pesquisa e descrevendo a origem geográfica das áreas científicas que levaram a esse desenvolvimento futuro.

2. REVISÃO DA LITERATURA

2.1. Inteligência Artificial

Em 1956, a Inteligência Artificial (IA) surgiu como uma disciplina científica devido à Conferência de *Dartmouth*, um evento histórico no qual o termo foi definido e uma agenda de pesquisa foi lançada para a criação de sistemas capazes de simular partes da inteligência humana. A IA também passou por fases desde então, de momentos em que tudo estava indo bem até os "invernos da IA", períodos de entusiasmo e estagnação.

Como explicaram Russell e Norvig (2016), a IA pode ser definida como a atividade de agentes racionais que percebem o ambiente e decidem como melhor maximizar seus objetivos. Isso enfatiza que este campo é interdisciplinar e inclui ciência da computação, matemática, psicologia cognitiva e filosofia. Nas décadas iniciais, a IA simbólica predominou—ela é baseada na modelagem de regras lógicas e sistemas especialistas. Modelos como *MYCIN e DENDRAL* exemplificam esta fase, que consiste na construção manual de regras de inferência. Esta abordagem, como Nilsson (2010) demonstrou, resultou em um alto grau de interpretabilidade, pois o raciocínio do sistema podia ser rastreado através das regras implementadas.

No entanto, a escalabilidade e adaptabilidade a contextos dinâmicos de tais sistemas eram limitadas. Ao longo dos anos 1990, ocorreu uma revolução significativa no paradigma da IA com a invenção de ferramentas estatísticas e aprendizado de máquina. A ênfase estava se afastando da codificação direta do que era conhecido, em direção ao aprendizado de dados. Como Bishop (2006) apontou, o aprendizado de máquina introduziu modelos probabilísticos e algoritmos capazes de generalizar padrões com base em grandes volumes de informação.

O impulso por mais generalização é ainda mais exacerbado pelo aumento do acesso a dados digitais e poder computacional, e, portanto, a proliferação de métodos algorítmicos à medida que se tornaram mais populares. O aprendizado profundo nasceu, houve uma nova era de evolução. LeCun, Bengio e Hinton (2015) ilustram que redes neurais profundas começaram a superar técnicas clássicas em tarefas complexas como reconhecimento de imagens e processamento de linguagem natural. Goodfellow, Bengio e Courville (2016) enfatizam a novidade fundamental do aprendizado profundo como o aprendizado automático de representações hierárquicas minimiza a engenharia manual de características. Esta trajetória demonstra a progressão da IA de sistemas baseados em regras transparentes para modelos complexos e orientados por dados.

Embora tenha sido um desenvolvimento significativo para o aumento de desempenho, isso levou a uma série de questões relacionadas à compreensão e interpretabilidade dos sistemas que foram criados e ao desenvolvimento adicional de IA explicável como uma solução para as limitações recentes do estado atual.

2.2. Desafios das IAs

A assimilação dos avanços recentes para a IA moderna introduziu uma série de desafios técnicos, éticos e sociais. Um dos principais problemas está relacionado à natureza de caixa preta dos modelos de aprendizado profundo. De fato, como Lipton (2016) argumenta, muitos dos modelos existentes são chamados de "caixas pretas" devido ao estado difícil de compreender de como cada parâmetro contribui para a escolha. Essa característica mina a capacidade de auditoria e dificulta a validação independente dos sistemas.

Outro desafio relevante está relacionado aos vieses algorítmicos. Sistemas treinados em dados históricos podem reproduzir desigualdades estruturais existentes na sociedade. Barocas e Selbst (2016) mostram que algoritmos podem produzir resultados discriminatórios mesmo quando não usam variáveis sensíveis explícitas (como raça ou gênero). O problema frequentemente é que há correlações indiretas nos dados. Isso é particularmente prejudicial na justiça criminal, crédito e recrutamento.

Outra questão importante é a confiabilidade e robustez dos sistemas. Os modelos de *deep learning* são vulneráveis a ataques adversários, como aqueles que ocorrem em cenários onde as perturbações dos dados de entrada podem causar grandes erros de classificação. Goodfellow, Shlens e Szegedy (2015) afirmam que essa observação é apoiada no estudo que mostrou que uma mudança em uma rede neural pode passar despercebida por humanos e isso questiona a segurança de aplicações sensíveis como veículos autônomos. Do ponto de vista regulatório, os avanços em IA tornam as questões de responsabilidade/transparência uma grande preocupação.

O Regulamento Geral sobre a Proteção de Dados (GDPR) da União Europeia gerou uma série de debates públicos sobre o direito à explicação para a tomada de decisões automatizadas. De acordo com a pesquisa de Wachter, Mittelstadt e Floridi (2017), o GDPR não estabelece um direito fundamental à explicação, mas implica a necessidade de habilitar medidas de transparência em sistemas automatizados. E a aceitação social da IA está enraizada na confiança do usuário. Doshi-Velez e Kim (2017) indicam que sistemas interpretáveis são cruciais para estabelecer confiança e habilitar o sistema em áreas importantes (saúde, finanças). Às vezes isso deve ser evitado, e essa resistência muitas vezes vem com a falta de explicações claras, especialmente quando decisões automatizadas são alcançadas de maneiras que resultam em impacto direto nas pessoas. Esses desafios para a IA não são puramente técnicos, mas também sociais, éticos e institucionais.

2.3. IA Explicável

A Inteligência Artificial Explicável (*XAI*) é uma resposta natural às limitações dos modelos opacos. A noção descreve a busca por maneiras de tornar os sistemas de IA mais fáceis de interpretar para os humanos, sem destruir drasticamente seus poderes de análise.

Arrieta *et al.,* (2020) definiram *XAI* como um conjunto de métodos que melhoram a transparência, interpretabilidade e compreensão dos processos de tomada de decisão automatizados.

O trabalho de Ribeiro, Singh e Guestrin (2016) está entre os marcos iniciais da *XAI* moderna, incluindo o LIME. Uma explicação local é produzida para qualquer modelo preditivo por este método. Sua contribuição original é que devemos permitir o entendimento da tomada de decisão aproximando o comportamento de um modelo a partir de um modelo linear simples e inferindo seus resultados localmente. Lundberg e Lee (2017) então propuseram o SHAP, um método derivado da teoria do valor de Shapley da teoria dos jogos cooperativos.

A abordagem usa o mesmo método acima e oferece consistência para medir o papel de cada variável na previsão do modelo. A popularização desses métodos trouxe a XAI para o âmbito da disciplina e da pesquisa estruturada. Conceitualmente falando, Molnar (2022) categoriza as abordagens de interpretabilidade em duas classes principais: algoritmos intrinsecamente interpretáveis (por exemplo, árvores de decisão, regressões lineares) e métodos pós-hoc adequados para modelos complexos. Acrescentando a este debate está Lipton (2016), que descreve a interpretabilidade como um conceito multidimensional, incluindo transparência estrutural e explicações decorrentes de aproximações externas. Além do fator técnico, a IA explicável tem um profundo impacto ético e regulatório.

Segundo Gunning (2017), o programa *XAI* da *DARPA* tinha como um dos principais objetivos aumentar a confiança do usuário no sistema avançado de IA. Sob este ângulo, a explicabilidade não é simplesmente um imperativo técnico, mas é vital para garantir a governança, a responsabilidade e a aceitação social da tecnologia. Assim, a *XAI* é uma inovação arquetípica que encapsula desempenho, transparência e responsabilidade dentro de uma estrutura científica idêntica.

3. MÉTODO

Este estudo caracteriza-se como uma pesquisa qualitativa-quantitativa com uma metodologia exploratória e descritiva, a fim de avaliar o desenvolvimento da Inteligência Artificial Explicável (*XAI*) nos níveis conceitual, técnico e geográfico.

O estudo foi desenvolvido ao longo de quatro dimensões complementares, incluindo revisão de literatura, análise bibliográfica, análise documental e análise bibliométrica.

3.1. Revisão de Literatura

O estudo realizou uma revisão de literatura para descobrir os principais marcos teóricos e metodológicos desde a Inteligência Artificial tradicional até a IA explicável. Isso envolveu a análise de obras clássicas no campo da IA, como Russell e Norvig (2016) e Goodfellow, Bengio e

Courville (2016), bem como literatura que consolidou o campo da *XAI*, como Ribeiro, Singh e Guestrin (2016), Lundberg e Lee (2017) e Arrieta *et al.,* (2020).

A estratégia de busca envolveu palavras-chave como “Inteligência Artificial”, “Aprendizado de Máquina”, “Aprendizado Profundo”, “Inteligência Artificial Explicável”, “Interpretabilidade” e “Transparência Algorítmica”. Esta seleção favoreceu artigos escritos em revistas indexadas, bem como conferências internacionais de alta qualidade no campo da ciência da computação e sistemas inteligentes.

Os artigos selecionados foram baseados em: (i) compreensão do entendimento histórico da evolução da IA, (ii) identificação de problemas técnicos e questões éticas relacionadas à obscuridade dos modelos, e (iii) a sistematização conceitual e técnica da IA explicável. O artigo é baseado em narrativa, onde paradigmas foram situados contextualmente no campo e lacunas na pesquisa existente foram identificadas.

3.2. Análise bibliográfica e documental

Os estudos bibliográficos consistem em ler e comparar os trabalhos acadêmicos selecionados para identificar quaisquer pontos potenciais de convergência, divergência e desenvolvimentos nas áreas de desenvolvimento de *XAI*. Foi um resumo não apenas dos aspectos conceituais da explicabilidade, das estratégias interpretáveis, das práticas e das discussões sobre ética e regulamentação, mas também da implementação. Também examinamos relatórios institucionais e recomendações regulatórias sobre governança de IA, além de produzirmos algumas análises documentais sobre esses assuntos.

Documentos primários (Gunning, 2017) do programa DARPA *XAI* e materiais relacionados ao Regulamento Geral de Proteção de Dados (GDPR) da União Europeia (Wachter, Mittelstadt, Floridi, 2017) foram identificados. Este exercício nos permitiu apreciar como a política pública e a implementação de medidas normativas internacionais contribuíram para o avanço atual nas agendas de explicabilidade. O foco da análise documental foi traçar um quadro mais claro do escopo e papel das agências governamentais, bem como das organizações internacionais na transparência algorítmica, e mapear os países com mais liderança na área de XAI.

3.3. Análise bibliométrica

A análise bibliométrica foi utilizada como técnica quantitativa na tentativa de analisar o desenvolvimento de textos científicos sobre Inteligência Artificial Explicável. Esta solução permite a comparação de publicações acadêmicas com base em parâmetros, por exemplo, o número de trabalhos publicados por ano, autores mais citados, países com maior volume de produção e principais áreas de aplicação.

As bases de dados científicas globais mais populares de tecnologia e engenharia neste campo foram levadas em consideração e observaram-se tendências temporais com base no crescimento do termo “Inteligência Artificial Explicável” a partir de 2016. A análise enfatizou a necessidade de identificar os principais países em termos de produção científica e concentrou-se nos Estados Unidos, China e países da União Europeia, que se sabe terem contribuído para a pesquisa em IA (Zhou *et al.,* 2019). As informações bibliométricas foram categorizadas tematicamente, para analisar a expansão da *XAI* em várias partes do mundo, incluindo saúde, finanças, direito e sistemas autônomos, respectivamente. A integração de métodos qualitativos (revisão e análise documental) e quantitativos (bibliometria) criou um quadro completo do avanço da IA explicável que delineou fundamentos teóricos, marcos estruturais e tendências internacionais na produção científica.

4. RESULTADOS

Esta seção reúne descobertas exclusivamente sobre Inteligência Artificial Explicável (*XAI*), com base na análise de dados bibliométricos de publicações indexadas usando os termos "*explainable AI" e "explainable artificial intelligence*", enfatizando o estudo de referência que analisou 4.781 documentos (Scopus, 2004–2023) (Sharma *et al.,* 2024).

4.1. Evolução da pesquisa sobre *XAI*

Os dados bibliométricos indicam que a *XAI* manteve baixa densidade até meados de 2017 e experimentou um crescimento acelerado a partir de 2018, consistente com a intensificação do debate sobre transparência/*interpretabilidade* e a popularização de métodos orientados para explicação e agendas de pesquisa (Sharma *et al.,* 2024). Como evidência, o estudo relata um aumento nas publicações de níveis residuais antes de 2017 para volumes significativamente mais altos no período de quatro anos seguinte, caracterizando uma fase de consolidação do campo no período de 2018–2022 (Sharma *et al.,* 2024).

Figura 1. Evolução das Pesquisas sobre XAI

Fonte: SHARMA *et al*., (2024).

4.2 Áreas de estudos sobre XAI

A análise de “estrutura intelectual” e concorrência de termos em *XAI* evidencia que o campo se organiza em torno de conceitos como explicabilidade, transparência, interpretação de

modelos e aprendizado profundo, além de ramificações para áreas aplicadas (Sharma *et al.,* 2024).

Tabela 1. Pesquisadores mais produtivos em XAI no corpus Scopus (2004–2023)

Autor	Nº de documentos	País(es)	Afiliação (conforme estudo)	h-index
Holzinger, A.	35	Áustria, Canadá	*Medical University Graz; Alberta Machine Intelligence Institute*	72
Guidotti, R.	22	Itália	*ISTI-CNR, Pisa*	23
Samek, W.	22	Alemanha	*Berlin Institute for the Foundations of Learning and Data*	57
Hagras, H.	18	Reino Unido	*University of Essex*	53
Alonso, J. M.	17	Espanha	*Universidade de Santiago de Compostela (CiTIUS)*	25
Främling, K.	16	Finlândia	*Aalto University*	36
André, E.	14	Alemanha	*University of Augsburg*	65
Biecek, P.	14	Polônia	*University of Warsaw*	32
Bobek, S.	14	Polônia	*Jagiellonian University (JAHCAI)*	15
Nalepa, G. J.	14	Polônia	*Jagiellonian University (JAHCAI)*	28

Fonte: Adaptado de Sharma et al. (2024).

4.3. Distribuição das pesquisas/consultas sobre XAI no mundo

No recorte bibliométrico do corpus analisado, a produção em *XAI* se concentra em polos tradicionais de pesquisa, com destaque para Estados Unidos e países europeus, além de crescimento relevante em países asiáticos (Sharma *et al.,* 2024).

Uma evidência complementar importante é que a XAI aparece fortemente vinculada a circuitos de publicação em conferências e séries de anais, além de periódicos e coleções editoriais técnicas (Sharma *et al.,* 2024).

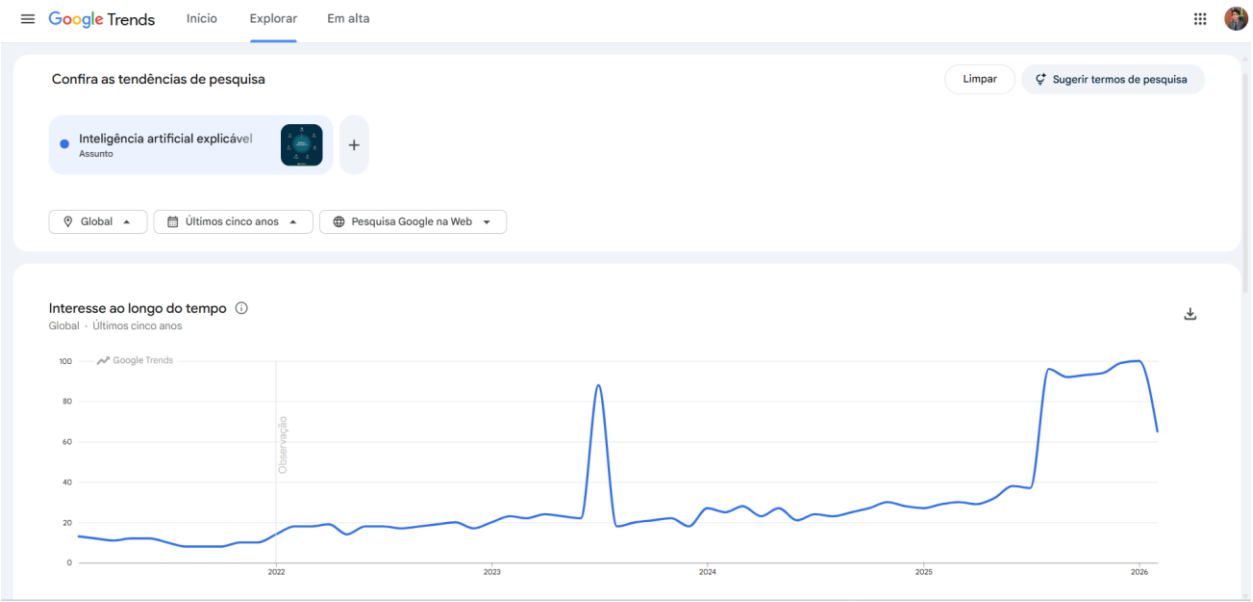
Tabela 2. Principais fontes/revistas e séries científicas com publicações em *XAI* (Source-wise) no *corpus* (2004–2023)

Fonte (revista / série / anais)	**Nº de publicações (documentos)**
Lecture Notes in Computer Science (LNCS)	497
CEUR Workshop Proceedings	243
IEEE Access	104
ACM International Conference (anais)	92
Communications in Computer and Information Science (CCIS)	82
Applied Sciences	70
IEEE International Conference on Fuzzy Systems (anais)	57
Sensors	51
Lecture Notes in Networks and Systems	50
Studies in Computational Intelligence	50

Fonte: adaptado de Sharma *et al.* (2024).

A seguir, complementa-se a distribuição geográfica da produção científica em XAI (seção 4.3 já apresentada com base em evidências bibliométricas) com um indicador de demanda social/informacional, isto é, o interesse de busca por XAI. Essa triangulação é coerente com o objetivo do artigo de entender a evolução da XAI não apenas como produção acadêmica, mas também como tema que ganha tração pública, influenciando adoção, regulação e prioridades de pesquisa (Sharma *et al.,* 2024).

Figura 2. Google Trends (interesse ao longo do tempo) para “Inteligência artificial explicável” (assunto), escopo global, últimos 5 anos, pesquisa na Web



Fonte: Google Trends (2026).

O gráfico do Google Trends mostra um índice de interesse relativo (0–100), com 100 como o pico de interesse do termo para a duração escolhida, e o restante dos valores representa uma relação proporcional a esse máximo (GOOGLE TRENDS, 2026). Entre 2021 e o início de 2022, o termo "inteligência artificial explicável" mostrou ter baixa penetração fora de nichos acadêmicos e técnicos, levando a um valor baixo e relativamente estável no início. No período de 2022 a meados de 2023, o interesse cresce gradualmente (em torno de ~10–15 para ~20–25) e isso pode ser entendido como um aumento orgânico na conscientização sobre esse campo, como visto em estudos bibliométricos realizados recentemente (Sharma *et al.*, 2024).

Isso é aparente pelo primeiro comportamento notável, um grande pico em 2023, que salta para um valor particularmente alto (próximo ao topo do gráfico) e rapidamente cai de volta ao seu valor anterior. Interpretado, este é o padrão típico para eventos "gatilho" — um evento de alto perfil (notícias, um lançamento tecnológico, um debate público, um curso massivo, um caso controverso sobre decisões automatizadas) que leva a buscas intensas por apenas um tempo relativamente curto. Como o Google Trends ainda não determinou o fenômeno causal de interesse até agora, a reação a este evento seria abordada com cautela: ele mostra um evento que recebeu atenção (um evento de choque), no entanto, você não pode fazer uma atribuição direta sem uma validação para notícias (e marcos regulatórios relevantes e cronogramas de lançamento) no mesmo período (GOOGLE TRENDS, 2026).

Este pico diz respeito à proposta do artigo porque destaca que, à medida que a *XAI* escala, ela não emerge apenas do acúmulo acadêmico, mas das tensões sociais do momento — fazendo-nos precisar mais de explicações e responsabilidade.

Esses dados sobem para uma faixa razoável após seu pico em 2023, e há algumas flutuações menores ao longo de 2024 com o gráfico permanecendo próximo do normal. Este fenômeno indica que o interesse não "relaxa" — ele se estabiliza em um segundo nível, implicando que a *XAI* continua a ser uma característica do discurso em torno de tópicos de consideração profissional e estudantil.

Essa é exatamente a estabilização (um novo normal, em níveis mais altos do que no início da série) que está relativamente presente em episódios que mudam de novidade episódica para competência desejável, e é representativa da proliferação da *XAI* como um requisito para avaliação de risco e certas aplicações sensíveis. É o próximo comportamento chave, de 2025 ao início de 2026: uma mudança clara de nível com o interesse atingindo níveis muito altos (em torno de ~90–100) e mantendo-se por um bom tempo. Ao contrário do pico de 2023, essa característica indica uma mudança sistêmica, não sazonal — então a duração permanece alta — semanas/meses.

Esta descoberta apoia ainda mais a interpretação de ver a questão da *XAI* como mais "institucionalizada" — ou seja, a explicabilidade é continuamente buscada, talvez ligada ao uso/adopção prática, conformidade, auditoria/exposição e propósito em produtos ou serviços. Embora não possamos tirar uma conclusão do gráfico e do estudo empírico, tais descobertas ecoam o que estudos bibliométricos revelam. Ou seja, a *XAI* está crescendo em volume e diversificando suas aplicações ao longo do tempo (Sharma *et al.,* 2024). Finalmente, a queda em seu último ponto (já em 2026) deve ser vista com cautela. No caso das séries do *Google Trends*, quedas à extrema direita ocorrem devido às mesmas razões, como uma janela de tempo incompleta (período ainda em consolidação) ou variações recentes no interesse.

Portanto, metodologicamente, é bom deixar o segmento final "provisório" e registrar no texto que é a data exata da coleta e o intervalo selecionado que determina a leitura do final da série (GOOGLE TRENDS, 2026).

De acordo com a proposta do estudo, essas descobertas apoiam duas conclusões sobre o tópico 4.3: (i) a liderança internacional de nações e entidades no processo de produção científica (avaliada via bibliometria) ocorre simultaneamente com uma dinâmica de interesse público global que pode se expandir com "ondas" e "saltos" relacionados a eventos e mudanças de adoção, e (ii) a *XAI* se solidifica como um tópico não apenas estudado nas universidades, mas também requisitado na sociedade, o que explica sua transição de um requisito "desejável" para um requisito "necessário" em aplicações de alto impacto. Complementando uma leitura de interesse ao longo do tempo representada anteriormente na Figura 2, a segunda evidência que extraímos

do *Google Trends* nos diz quais termos as pessoas realmente pesquisam quando "inteligência artificial explicável" é o tópico, separando consultas mais frequentes (de volume estável) de consultas em ascensão (crescimento recente). Essa camada é vital para o objetivo do nosso artigo, pois permite insights sobre por que o interesse aumenta ou diminui e questões sobre os múltiplos significados do termo "*XAI*" no domínio público — o que, por sua vez, informa nossa análise da "distribuição de consultas sobre *XAI* no mundo" (GOOGLE TRENDS, 2026).

Considerando as consultas típicas na investigação atual, é possível inferir que os termos mais perguntados aqui são altamente amplos e conceituais: "IA", "explicável", "IA explicável", "explicabilidade" e assim por diante. Esse padrão sugere que a grande maioria do público pode estar sendo introduzida ao tema em sua variedade genérica (IA) ou exploratória (explicabilidade) em sua experiência diária, e que os dois tipos de dados estão alinhados com a discussão sobre transparência e interpretabilidade quando aplicados no uso contínuo (Sharma *et al.,* 2024).

Embora seja aí que "*XAI*" aparece ou não de fato nessas contagens, sua diminuição relativa na lista é significativamente diferente da alta variância de "explicabilidade", assim como tipos com um crescimento significativo semelhante às combinações como "IA explicável" e "explicabilidade da IA", que também mostram um aumento significativo (por exemplo, +200%−+250% na captura).

Interpretativamente, isso significa que o público quer mais abstração ou elucidação do que um acrônimo de *XAI* e isso se torna mais evidente quando o acrônimo começa a ser confundido com outras definições além de algumas acadêmicas (GOOGLE TRENDS, 2026). A gama de termos é ainda mais clara ao ver a coluna de consultas ascendentes; termos como "OpenAI", "*XAI* Musk", "Elon Musk *XAI*", "*XAI* crypto", "*XAI* token", "*XAI* coin", "Grok" e "site *XAI*" estão fortemente salpicados com "Grande Aumento". Essa sequência de blocos prova que os picos de buscas por 'XAI' no período recente não são tanto na própria XAI (a nova palavra) quanto na marca/empresa *XAI* e tópicos vinculados (produto, notícias, ativos de criptomoeda), especialmente no que se refere ao termo acadêmico *XAI* como Inteligência Artificial Explicável. Tradução: o termo "XAI" é semanticamente ambíguo e não contém contexto quando usado no domínio público, o que pode levar a medições de interesse infladas ou distorcidas se a coleta descrever apenas *XAI* no domínio público (GOOGLE TRENDS, 2026).

Esta observação pode explicar — e pode estar relacionada a — por que o interesse pode ser mostrado como saltos em um gráfico de linha do tempo devido ao fato de que os picos não são tanto resultado de interesse em "explicabilidade" quanto resultado de eventos ou atenção da mídia para "*XAI*" (marca).

Do ponto de vista da pesquisa, também poderíamos concluir que, para o artigo, é mais apropriado ler cuidadosamente a "distribuição de consultas", onde o *Google Trends* pode indicar a demanda social geral e a liderança científica de acordo com o país, onde a influência científica

provavelmente deve ser presumida com base em informações bibliométricas (publicações, citações, redes) relatadas a partir de uma análise específica de XAI (Sharma *et al.,* 2024).

Metodologicamente, também demonstra que os achados podem ser ainda mais refinados: em pontos onde o objetivo é despertar interesse em IA Explicável (área científica), sugeriríamos classificar expressões menos ambíguas (por exemplo, IA explicável, explicabilidade ou aprendizado de máquina explicável) e/ou usar combinações/segmentações para minimizar o ruído (por exemplo, comparar "IA explicável" com "XAI" e relatar resultados diferentes).

Portanto, o estudo apresenta consistência entre o propósito do indicador (mostrar interesse em XAI acadêmico) e o verdadeiro valor medido (interesse acadêmico vs interesse convergente de marca/evento) (GOOGLE TRENDS, 2026). Assim, a seção 4.3 reúne ambas as evidências, a saber, uma linha do tempo e consultas relacionadas, chegando à conclusão de que XAI ainda é um tema comum de pesquisa, mas o aumento nos últimos anos pode ser atribuído à competição associada às definições de XAI. Isso não dilui o sinal; ao contrário, amplia seu alcance: apresenta a agenda de explicabilidade como crescendo junto com a "popularização" do vocabulário de IA — portanto, sendo cauteloso em relação aos termos entre países, tendências e épocas (Sharma *et al.,* 2024; GOOGLE TRENDS, 2026).

CONSIDERAÇÕES

Essa pesquisa indica que a transição da IA tradicional para a *XAI* está sendo impulsionada não apenas pela inovação algorítmica avançada, mas pela interseção de múltiplos fatores: primeiro, modelos mais complexos e opacos; segundo aplicações de alto impacto mais concentradas; terceiro, preocupações com preconceitos e justiça algorítmica; e quarto, pressões regulatórias e institucionais por transparência e responsabilidade.

A análise bibliométrica demonstra o surgimento, como agenda científica, da *XAI*, em particular a aceleração de um corpo de publicações desde 2018 no espaço e a posição proeminente que autores específicos e fontes editoriais ocupam neste campo. Além disso, os resultados do *Google Trends* sugerem que há uma demanda social crescente, mas também apontam uma importante questão metodológica: ao buscar informações, o termo "*XAI*" pode ser usado para significados não acadêmicos, então deve-se evitar o uso de terminologia enganosa (como "IA explicável" e "explicabilidade") na esperança de quantificar o interesse público no conceito científico em questão.

Além da evolução técnica, observa-se que a consolidação da Inteligência Artificial Explicável se insere em um movimento mais amplo de reorientação da própria agenda da inteligência artificial para um modelo centrado no ser humano (*human-centered AI*). A incorporação de princípios como transparência, responsabilidade e mitigação de vieses não representa apenas uma exigência regulatória, mas um reconhecimento de que sistemas

algorítmicos passam a interagir diretamente com dimensões tradicionalmente associadas à racionalidade e à tomada de decisão humanas (COMISSÃO EUROPEIA, 2019; Wachter; Mittelstadt; Floridi, 2017). Nesse contexto, a XAI pode ser compreendida como tentativa de reduzir a assimetria cognitiva entre modelos computacionais altamente complexos e seus usuários, buscando preservar autonomia, compreensão e capacidade crítica diante de decisões automatizadas. Tal movimento evidencia que o avanço tecnológico contemporâneo não se limita ao aumento de desempenho preditivo, mas envolve também a incorporação de princípios normativos que historicamente estruturam a convivência social.

As conclusões bibliométricas dependem do banco de dados, período e estratégia de busca e, portanto, são limitações, enquanto os indicadores de busca indicam interesse relativo e, portanto, podem ser vítimas de ruído semântico. Para a agenda futura, propomos expandir nossa bibliometria com redes de coautoria e cocitação e tendências comparativas por campos aplicados (saúde, finanças, setor público) que melhorariam ainda mais a integração de métricas/padrões/guias/ferramentas de adoção, o que pode fornecer um mapa melhor de como a explicabilidade está mudando da pesquisa para o *design* e governança da IA.

REFERÊNCIAS

ALVES, Davis Souza; LIMA, Adrianne Correia. **Encarregados**: Data Protection Officer (DPO). [S. l.]: Editora Haikai, 2021.

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.

BAROCAS, Solon; SELBST, Andrew D. Big data's disparate impact. **California Law Review**, Berkeley, v. 104, n. 3, p. 671–732, 2016.

BISHOP, Christopher M. **Pattern recognition and machine learning**. New York: Springer, 2006.

COMISSÃO EUROPEIA. **Ethics guidelines for trustworthy AI**. Bruxelas: European Commission, 2019.

DOSHI-VELEZ, Finale; KIM, Been. Towards a rigorous science of interpretable machine learning. **arXiv preprint**, arXiv:1702.08608, 2017.

GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. Cambridge: MIT Press, 2016.

GOODFELLOW, Ian; SHLENS, Jonathon; SZEGEDY, Christian. Explaining and harnessing adversarial examples. **arXiv preprint**, 2015.

GOOGLE TRENDS. **Inteligência artificial explicável (assunto)**. Escopo: global; intervalo: últimos cinco anos; tipo: Pesquisa Google na Web. Acesso em: 18 fev. 2026.

GOV.UK. **Data ethics framework**. London: Government Digital Service, 2018. Atualizado em 2025.

GUNNING, David. **Explainable artificial intelligence (XAI)**. Arlington: DARPA, 2017.

ICO – INFORMATION COMMISSIONER'S OFFICE. **Explaining decisions made with AI**. London: ICO, 2020.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, London, v. 521, p. 436–444, 2015.

LIPTON, Zachary C. The mythos of model interpretability. **Communications of the ACM**, v. 61, n. 10, p. 36–43, 2018.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. *In:* **Advances in Neural Information Processing Systems (NeurIPS)**, 2017.

MOLNAR, Christoph. **Interpretable machine learning**. 2. ed. [S. l.]: Leanpub, 2022.

NILSSON, Nils J. **The quest for artificial intelligence**: a history of ideas and achievements. Cambridge: Cambridge University Press, 2010.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "Why should I trust you?" Explaining the predictions of any classifier. *In:* **Proceedings** […] of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016. p. 1135–1144.

RUSSELL, Stuart; NORVIG, Peter. **Artificial intelligence**: a modern approach. 3. ed. Upper Saddle River: Pearson, 2016.

SHARMA, Chetan et al. Exploring explainable AI: a bibliometric analysis. **Discover Applied Sciences**, v. 6, 2024.

WACHTER, Sandra; MITTELSTADT, Brent; FLORIDI, Luciano. Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. **International Data Privacy Law**, Oxford, v. 7, n. 2, p. 76–99, 2017.

ZHOU, Zhi-Hua et al. Machine learning in China: past, present and future. **National Science Review**, Oxford, v. 6, n. 1, p. 19–38, 2019.