REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

THE EVOLUTION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE (XAI)

A EVOLUÇÃO DA INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (IAE)

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

PUBLISHED: 03/2026

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

[1] Master's student in Administration with research in Explainable Artificial Intelligence (XAI), postgraduate in IT Management and Cloud Computing (UFSCar) and graduate in IT Management (UNIP), with international certifications in ITIL® 4, Security+, DPO and ISO 27001. Mid-level Support Analyst at Stefanini Brazil (BAT Latam South) and IT Junior Consultant. Serves as an academic evaluator in IT programs and National Coordinator of the APDADOS Startech Committee.

[2] Master's degree in Administration from Florida Christian University with research in Advanced Emerging Technologies. MBA in IT Project Management and international certifications such as Certified Information Security Officer (CISO), ISO/IEC 27001 Lead Implementer and DPO – EXIN. Bachelor's degree in Computer Network Technology. Extensive experience leading projects in companies such as NTT Ltd., ISH Tecnologia and Grupo Cornélio Brennand.

[3] Technologist in Marketing (UNIP) and postgraduate in Data Protection Officer (LGPD/GDPR), with specialization in Social Media and experience managing social networks for IT professionals and third-sector institutions. Marketing Manager and Advisor to the APDADOS Steering Committee, with institutional work in Brasília with federal agencies between 2021 and 2023. Internationally, participated in official missions to countries such as France, England and Angola, where served as an invited speaker by the Minister of Technology.

[4] PhD in IT Administration from Florida Christian University (USA), validated in Brazil; Master's degree in Administration with a focus on Green IT (2015); Extension in IT Management from FGV/SP (2011); Postgraduate in Project Management (2009). Professor of Information Security at Universidade Paulista (UNIP), Universidade Municipal de São Caetano do Sul (USCS) and Florida Christian University (FCU). Holds PMP®, ITIL® Expert, C|EH®, C|HFI® and EXIN Data Protection certifications. Works in the United States as a Cybersecurity Project Manager (R&D) focusing on Data Privacy (LGPD/GDPR), Digital Forensics, Ethical Hacking and Artificial Intelligence (AI). President of the National Association of Data Privacy Professionals (APDADOS.org).

[5] PhD in Economics from PUC-Campinas. MBA in Marketing – ESAMC (Sorocaba). Master's degree in Administration from Universidade Guarulhos (UNG). Master's degree in Sociology from PUC-São Paulo. PhD in Sociology from PUC-São Paulo. PhD in Administration from FCU-USA. Postdoctoral studies at UNICAMP (Campinas), FCU-USA and University of Coimbra (Portugal). Journalist and Writer. Evaluator for MEC/INEP. Vice-Rector of Universidade Guarulhos, SP.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

evidence (Scopus, 2004–2023) indicate rapid growth of XAI publications from 2018 onwards, 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 analyzes 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. Model Interpretability. Algorithmic Transparency. Post-hoc Explanations. Deep Learning. AI Governance. Algorithmic Auditing. Algorithmic Bias. Bibliometrics.

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 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.

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

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

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. INTRODUCTION

Artificial intelligence (AI) has come a long way since it was developed as a scientific field in the 1950s. It has been derived from symbolic approaches and rule-based systems, but it was "traditional AI," with its explicit representation of field knowledge, that made its decision-making processes relatively transparent. Structured models such as expert systems and decision trees allow for a structured exploration of rules and inferences (Russell; Norvig, 2016). However, along with the progressive improvement of statistical methods and machine learning, most notably since the 2000s, the field has witnessed a paradigm shift from symbolic models to data-driven models – leading to the rise of machine learning and deep learning.

There were three main reasons for the growth of deep learning: improved computing power, large data availability, and improvements in machine learning algorithms in artificial neural networks (Lecun; Bengio; Hinton, 2015). For example, these factors led to a significant improvement in image recognition tasks, natural language processing, and recommendation systems. However, the greater structural complexity of these models was accompanied by a major weakness: the black box, or opacity, of these systems (Goodfellow; Bengio; Courville, 2016).

While in conventional symbolic systems it was possible to see where a conclusion had been reached from the logical sequence of events, models that use deep neural networks have millions or even billions of parameters that are automatically adjusted from data. This characteristic leads to a challenge for humans in interpreting the results of their decision-making. As Lipton (2016) argues, the meaning of " interpretability " has gained wide popularity, but it has also been commonly interpreted ambiguously, including: structural transparency, post-hoc explanation, cognitive simulation capability.

The increasing use of AI in urgent circumstances has made explainability even more necessary. Automated systems have begun to appear in medical diagnoses, credit granting, hiring, legal matters, and public safety. In these contexts, automated decisions can impact the rights of individuals or even communities, and therefore it is critical to understand how and why a specific decision was reached (Doshi-Velez; Kim, 2017). A lack of explanations can undermine trust, accountability, and social legitimacy regarding these technologies.

Furthermore, the literature indicates the ability of AI models to emulate or amplify biases embedded in the training dataset. Barocas and Selbst (2016) show that algorithmic systems can have discriminatory effects without intentionally creating them and can, in fact, reinforce structural

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

inequality. This case reinforced the demand for an audit mechanism and interpretations of automated decisions, for greater transparency, and for fairness in our algorithms.

In this sense, Explainable Artificial Intelligence (XAI) becomes evident and establishes itself as a subfield for developing and building techniques to understand complex models. The term gained worldwide acceptance starting in 2016, particularly through DARPA's XAI program (Gunning, 2017), and its focus on creating systems that would explain information to human operators.

Several scientific advances have facilitated the consolidation of explainable AI. To this end, Ribeiro *et al.,* (2016) presented the LIME (Local Interpretable Model- Agnostic) method. Explanations), which allows explanations for individual decisions to be given in reasonable local estimates by the respective classifier. Lundberg and Lee (2017) then introduced SHAP (SHapley). Additive exPlanations) to develop an adaptive framework that can be used to consider unique assignments of variable importance in complex models, based on cooperative game theory.

Furthermore, as suggested by Molnar (2022), the authors made a logical typology of interpretability methods between global and local, intrinsic and post-hoc explanations, thus providing a conceptual basis for research. Arrieta *et al.,* (2020) offer an extensive literature review on the topic of XAI, which examines concepts including interpretability, explainability, transparency, comprehensibility, and the broad interdisciplinary aspects of the field.

The development of explainable AI has also been informed by political debates. The European Union's General Data Protection Regulation (GDPR), issued in 2018, stimulated discussions about the so-called "right to explanation" regarding automated decisions, and it is up to Data Protection Officers (DPOs) to encourage these discussions within companies (Alves, 2021). Although there are controversial legal interpretations of this right, Wachter, Mittelstadt, and Floridi (2017) consider the legality of algorithmic transparency and the need for explanatory mechanisms to understand the complexity of the systems.

Thus, the shift from traditional AI to explainable AI can be interpreted as a composite of four intersecting axes (i) technical sophistication of models, (ii) increasing use in sensitive contexts, (iii) ethical considerations regarding bias and discrimination, and (iv) regulatory challenges surrounding transparency and accountability. Some countries in the world, for example, are at the forefront of scientific advancements in both AI and XAI.

Based on bibliometric data, the United States leads the world in the number of publications on artificial intelligence, with universities such as MIT, Stanford, and Carnegie Mellon leading this growth. AI has seen remarkably rapid scientific output in China over the decades and is now becoming a key global hub (Zhou *et al.,* 2019).

In Europe, the United Kingdom, Germany, and France are at the forefront of research that harmonizes artificial intelligence, ethics, and technological regulation.

ISSN: 2675-6218 - RECIMA21

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

For example, the United Kingdom is well-known for articulating the dialogue between academia and public policy when it comes to algorithmic governance. Thus, the establishment of explainable AI is a step not only in the technical direction but also in relation to the social demand for transparency, reliability, and accountability of automated systems. It is a fundamental paradigm shift that places a greater focus on understandability and accountability as central demands beyond simply predicting with the new model.

The research question was developed under these conditions: What factors contributed to the evolution from Conventional Artificial Intelligence to Explainable Artificial Intelligence, and which countries are prominent in this regard?

In response to this question, this article aims to study how explainable AI has developed over the years, focusing on the main theoretical and methodological milestones of the research and describing the geographical origin of the scientific areas that led to this future development.

2. LITERATURE REVIEW

2.1. Artificial Intelligence

First, in 1956, Artificial Intelligence (AI) emerged as a scientific discipline due to the Dartmouth Conference, a landmark event in which the term was defined, and a research agenda was launched for the creation of systems capable of simulating aspects of human intelligence. AI has also gone through phases since then, from periods of great success to "AI winters," periods of both enthusiasm and stagnation.

As Russell and Norvig (2016) explained, AI can be defined as the activity of rational agents that perceive the environment and decide how best to maximize their objectives. This emphasizes that this field is interdisciplinary and includes computer science, mathematics, cognitive psychology, and philosophy. In the early decades, symbolic AI predominated - it is based on the modeling of logical rules and expert systems. Models such as MYCIN and DENDRAL exemplify this phase, which consists of the manual construction of inference rules. This approach, as Nilsson (2010) demonstrated, resulted in a high degree of interpretability, since the system's reasoning could be traced through the implemented rules.

However, the scalability and adaptability to dynamic contexts of such systems were limited. Throughout the 1990s, a significant revolution occurred in the AI paradigm with the invention of statistical tools and machine learning. The emphasis was shifting away from directly encoding what was known, towards learning from data. As Bishop (2006) pointed out, machine learning introduced probabilistic models and algorithms capable of generalizing patterns based on large volumes of information.

The drive for greater generalization is further exacerbated by increased access to digital data and computing power, and therefore the proliferation of algorithmic methods as they have

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

become more popular. Deep learning was born, and a new era of evolution began. LeCun, Bengio and Hinton (2015) illustrate that deep neural networks have begun to outperform classical techniques in complex tasks such as image recognition and natural language processing. Goodfellow, Bengio and Courville (2016) emphasize the fundamental novelty of deep learning as the machine learning of hierarchical representations minimizes manual feature engineering. This trajectory demonstrates the progression of AI from transparent rule-based systems to complex, data-driven models.

Although this was a significant development for increasing performance, it led to a number of issues related to the understanding and interpretability of the systems that were created, and to the further development of explainable AI as a solution to the recent limitations of the current state.

2.2. Challenges of AI

The assimilation of recent advances in modern AI has introduced a number of technical, ethical, and social challenges. One of the main problems is related to the black-box nature of deep learning models. In fact, as Lipton (2016) argues, many existing models are called "black boxes" due to the difficult-to-understand state of how each parameter contributes to the choice. This characteristic undermines auditability and hinders independent validation of the systems.

Another relevant challenge relates to algorithmic biases. Systems trained on historical data can reproduce existing structural inequalities in society. Barocas and Selbst (2016) show that algorithms can produce discriminatory results even when they do not use explicit sensitive variables (such as race or gender). The problem is often that there are indirect correlations in the data. This is particularly harmful in criminal justice, credit, and recruitment.

Another important issue is the reliability and robustness of the systems. Deep learning models are vulnerable to adversarial attacks, such as those that occur in scenarios where perturbations in input data can cause large classification errors. Goodfellow, Shlens and Szegedy (2015) state that this observation is supported by a study showing that a change in a neural network can go unnoticed by humans, raising questions about the security of sensitive applications such as autonomous vehicles. From a regulatory standpoint, advances in AI make accountability/transparency issues a major concern.

The European Union's General Data Protection Regulation (GDPR) has generated a series of public debates about the right to explanation for automated decision-making. According to research by Wachter, Mittelstadt and Floridi (2017), the GDPR does not establish a fundamental right to explanation, but it does imply the need to enable transparency measures in automated systems. And the social acceptance of AI is, of course, rooted in user trust. Doshi-Velez and Kim (2017) indicate that interpretable systems are crucial for establishing trust and enabling the system in important areas (healthcare, finance).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Sometimes this must be avoided, and this resistance often stems from a lack of clear explanations, especially when automated decisions are reached in ways that directly impact people. These challenges for AI are not purely technical, but also social, ethical, and institutional.

2.3. Explainable AI

Explainable Artificial Intelligence (XAI) is a natural response to the limitations of opaque models. The concept describes the search for ways to make AI systems easier for humans to interpret without drastically destroying their analytical powers. Arrieta *et al.,* (2020) defined XAI as a set of methods that improve the transparency, interpretability and understanding of automated decision-making processes.

The work of Ribeiro, Singh, and Guestrin (2016) is among the initial milestones of modern XAI, including LIME. A local explanation is produced for any predictive model by this method. Their original contribution is that we should allow the understanding of decision-making by approximating the behavior of a model from a simple linear model and inferring its results locally. Lundberg and Lee (2017) then proposed SHAP, a method derived from Shapley 's value theory of cooperative game theory.

The approach uses the same method as above and offers consistency in measuring the role of each variable in the model's prediction. The popularization of these methods has brought XAI into the realm of discipline and structured research. Conceptually speaking, Molnar (2022) categorizes interpretability approaches into two main classes: intrinsically interpretable algorithms (e.g., decision trees, linear regressions) and post-hoc methods suitable for complex models. Adding to this debate is Lipton (2016), who describes interpretability as a multidimensional concept, including structural transparency and explanations arising from external approximations. Beyond the technical factor, explainable AI has a profound ethical and regulatory impact.

According to Gunning (2017), one of the main objectives of DARPA's XAI program was to increase user confidence in advanced AI systems. From this perspective, explainability is not simply a technical imperative, but is vital to ensuring governance, accountability, and social acceptance of the technology. Thus, XAI is an archetypal innovation that encapsulates performance, transparency, and accountability within an identical scientific framework.

3. METHODOLOGY

This study was distinguished as a qualitative and quantitative research with an exploratory and descriptive methodology, in order to evaluate the development of Explainable Artificial Intelligence (XAI) at the conceptual, technical and geographical levels.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

The study was developed along four complementary dimensions, including literature review, bibliographic analysis, document analysis, and bibliometric analysis.

3.1. Literature Review

The study conducted a literature review to uncover the main theoretical and methodological milestones from traditional Artificial Intelligence to explainable AI. This involved analyzing classic works in the field of AI, such as Russell and Norvig (2016) and Goodfellow, Bengio and Courville (2016), as well as literature that consolidated the field of XAI, such as Ribeiro, Singh, and Guestrin (2016), Lundberg and Lee (2017), and Arrieta *et al.,* (2020).

The search strategy involved keywords such as “Artificial Intelligence”, “Machine Learning”, “Deep Learning”, “Explainable Artificial Intelligence”, “Interpretability”, and “Algorithmic Transparency”. This selection favored articles published in indexed journals, as well as high-quality international conferences in the field of computer science and intelligent systems.

The selected articles were based on: (i) understanding the historical understanding of the evolution of AI, (ii) identifying technical problems and ethical issues related to the obscurity of the models, and (iii) the conceptual and technical systematization of explainable AI. The article is narrative-based, where paradigms were contextually situated within the field and gaps in existing research were identified.

3.2. Bibliographic and Documentary Analysis

Bibliographic studies consist of reading and comparing selected academic works to identify any potential points of convergence, divergence, and developments in the areas of XAI development. It was a summary not only of the conceptual aspects of explainability, interpretable strategies, practices, and discussions on ethics and regulation, but also of implementation. We also examined institutional reports and regulatory recommendations on AI governance, in addition to producing some documentary analyses on these subjects.

Primary documents (Gunning, 2017) from the DARPA XAI program and materials related to the European Union's General Data Protection Regulation (GDPR) (Wachter, Mittelstadt, Floridi, 2017) were identified. This exercise allowed us to appreciate how public policy and the implementation of international normative measures have contributed to the current progress on explainability agendas. The focus of the document analysis was to draw a clearer picture of the scope and role of government agencies, as well as international organizations, in algorithmic transparency, and to map the countries with the most leadership in the area of XAI.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

3.3. Bibliometric Analysis

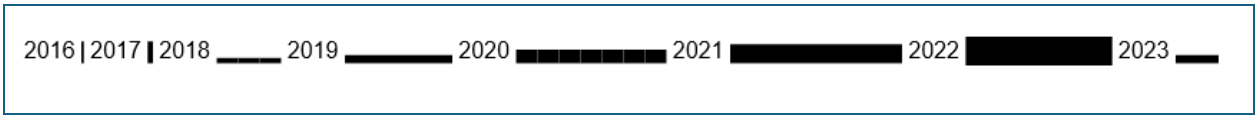
Bibliometric analysis was used as a quantitative technique in an attempt to analyze the development of scientific texts on Explainable Artificial Intelligence. This solution allows for the comparison of academic publications based on parameters such as the number of papers published per year, most cited authors, countries with the highest volume of production, and main areas of application. The most popular global scientific databases in technology and engineering in this field were considered, and we were able to observe temporal trends based on the growth of the term "Explainable Artificial Intelligence" from 2016 onwards. The analysis emphasized the need to identify the main countries in terms of scientific production and focused on the United States, China, and countries of the European Union, which are known to have contributed to AI research (Zhou *et al.,* 2019). The bibliometric information was thematically categorized to analyze the expansion of XAI in various parts of the world, including health, finance, law, and autonomous systems, respectively. The integration of qualitative (review and document analysis) and quantitative (bibliometrics) methods created a comprehensive picture of the advancement of explainable AI, outlining theoretical foundations, structural frameworks, and international trends in scientific production.

4. RESULTS (re-presentation focusing exclusively on XAI)

This section brings together findings exclusively on Explainable Artificial Intelligence (XAI), based on the analysis of bibliometric data from publications indexed using the terms " explainable AI" and " explainable artificial intelligence," emphasizing the landmark study that analyzed 4,781 documents (Scopus, 2004–2023) (Sharma *et al.,* 2024). 4.1 Evolution of research on XAI

Bibliometric data indicate that XAI maintained low density until mid-2017 and experienced accelerated growth from 2018 onwards, consistent with the intensification of the debate on transparency/ interpretability and the popularization of explanation-oriented methods and research agendas (Sharma *et al.,* 2024). As evidence, the study reports an increase in publications from residual levels before 2017 to significantly higher volumes in the following four-year period, characterizing a consolidation phase of the field in the period 2018–2022 (Sharma *et al.,* 2024).

Figure 1. Evolution of Research on XAI



Source: SHARMA et al., (2024).

ISSN: 2675-6218 - RECIMA21
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

4.1. Areas of study on XAI

The analysis of "intellectual structure" and co-occurrence of terms in XAI shows that the field is organized around concepts such as explainability, transparency, model interpretation, and deep learning, in addition to ramifications for applied areas (Sharma *et al.,* 2024).

Table 1. Most productive researchers in XAI in the Scopus corpus (2004–2023)

Author	Number of documents	Country(ies)	Affiliation (as per study)	h-index
Holzinger, A.	35	Austria, Canada	Medical University Graz; Alberta Machine Intelligence Institute	72
Guidotti, R.	22	Italy	ISTI-CNR, Pisa	23
Samek, W.	22	Germany	Berlin Institute for the Foundations of Learning and Data	57
Hagras , H.	18	United Kingdom	University of Essex	53
Alonso, JM	17	Spain	University of Santiago de Compostela (CiTIUS)	25
Främling , K.	16	Finland	Aalto University	36
André, E.	14	Germany	University of Augsburg	65
Biecek, P.	14	Poland	University of Warsaw	32
Bobek, S.	14	Poland	Jagiellonian University (JAHCAI)	15
Nalepa, GJ	14	Poland	Jagiellonian University (JAHCAI)	28

Source: Adapted from Sharma et al. (2024).

4.2. Distribution of XAI research/consultations worldwide

In the bibliometric analysis of the corpus examined, research in XAI is concentrated in traditional research centers, particularly the United States and European countries, in addition to significant growth in Asian countries (Sharma *et al.,* 2024).

An important complementary piece of evidence is that XAI appears strongly linked to publication circuits in conferences and proceedings series, as well as technical journals and editorial collections (Sharma *et al.,* 2024).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

Table 2. Main scientific sources/journals and series with XAI (Source-wise) publications in the corpus (2004–2023)

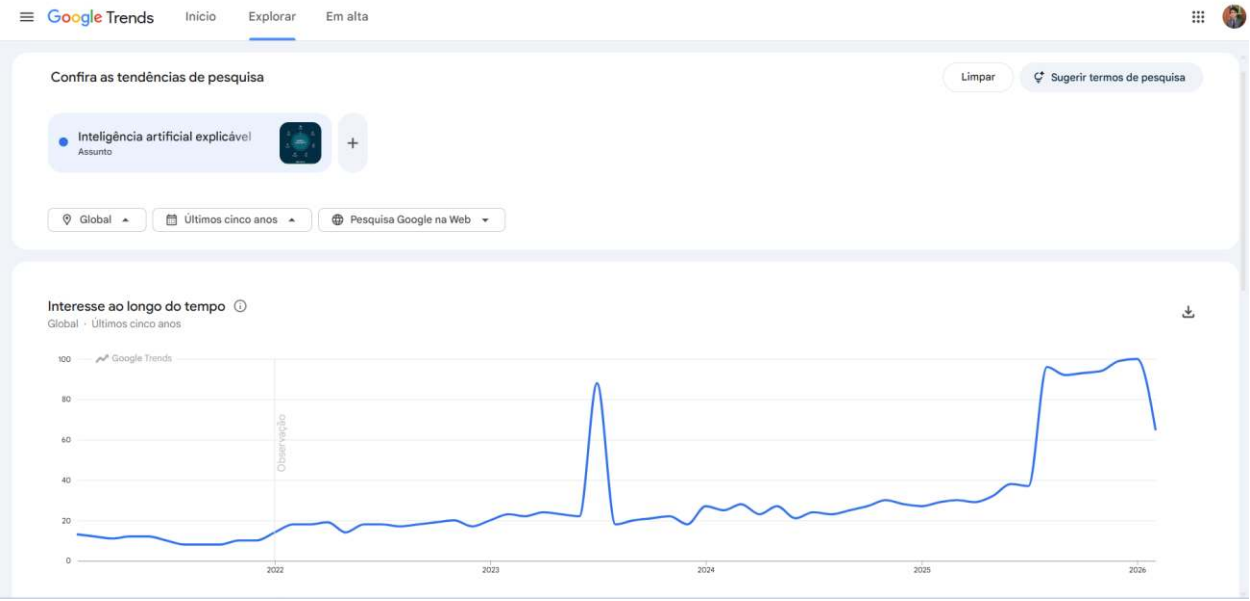
Source (journal / series / proceedings)	Number of publications (documents)
Lecture Notes in Computer Science (LNCS)	497
CEUR Workshop Proceedings	243
IEEE Access	104
ACM International Conference (Proceedings)	92
Communications in Computer and Information Science (CCIS)	82
Applied Sciences	70
IEEE International Conference on Fuzzy Systems (Proceedings)	57
Sensors	51
Lecture Notes in Networks and Systems	50
Studies in Computational Intelligence	50

Source: adapted from Sharma et al. (2024).

Next, the geographical distribution of scientific production in XAI (section 4.3 already presented based on bibliometric evidence) is complemented by an indicator of social/informational demand, that is, the search interest in XAI. This triangulation is consistent with the article's objective of understanding the evolution of XAI not only as academic production, but also as a topic that gains public traction, influencing adoption, regulation, and research priorities (Sharma *et al.,* 2024).

ISSN: 2675-6218 - RECIMA21
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

Figure 2. Google Trends (interest over time) for “Explainable Artificial Intelligence” (topic), global scope, last 5 years, web search.



Source: Google Trends (2026).

The Google Trends graph shows a relative interest index (0–100), with 100 representing the peak interest in the term for the chosen duration, and the remaining values representing a proportional relationship to this maximum (GOOGLE TRENDS, 2026). Between 2021 and early 2022, the term "explainable artificial intelligence" showed low penetration outside of academic and technical niches, leading to a low and relatively stable value initially. In the period from 2022 to mid-2023, interest gradually increased (around ~10–15 to ~20–25), and this can be understood as an organic increase in awareness of this field, as seen in recently conducted bibliometric studies (Sharma *et al.,* 2024).

This is apparent from the first notable behavior, a large spike in 2023, which jumps to a particularly high value (near the top of the graph) and quickly falls back to its previous value. Interpreted, this is the typical pattern for "trigger" events—a high-profile event (news, a technology launch, a public debate, a massive course, a controversial case about automated decisions) that leads to intense searches for only a relatively short time. As Google Trends has not yet determined the causal phenomenon of interest so far, the reaction to this event should be approached with caution: it shows an event that received attention (a shock event), however, you cannot make a direct attribution without validation to news (and relevant regulatory milestones and launch schedules) in the same period (GOOGLE TRENDS, 2026).

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

This peak relates to the article's proposal because it highlights that, as XAI escalates, it emerges not only from academic accumulation but also from the social tensions of the moment—making us need more explanations and accountability.

These figures rise to a reasonable range after their peak in 2023, and there are some minor fluctuations throughout 2024 with the graph remaining close to normal. This phenomenon indicates that interest does not "relax"—it stabilizes at a second level, implying that XAI continues to be a feature of the discourse surrounding topics of professional and student consideration.

This is precisely the stabilization (a new normal, at higher levels than at the beginning of the series) that is relatively present in episodes that shift from episodic novelty to desirable competence and is representative of the proliferation of XAI as a requirement for risk assessment and certain sensitive applications. The next key behavior, from 2025 to early 2026, is a clear level change with interest reaching very high levels (around ~90–100) and remaining there for a considerable time. Unlike the 2023 peak, this characteristic indicates a systemic, not seasonal, change—hence the duration remains high—weeks/months.

This finding further supports the interpretation of viewing the issue of XAI as more "institutionalized"—that is, explainability is continuously sought, perhaps linked to practical use/adoption, compliance, audit/exposure, and purpose in products or services. While we cannot draw a conclusion from the graph and empirical study, these findings echo what bibliometric studies reveal. That is, XAI is growing in volume and diversifying its applications over time (Sharma *et al.,* 2024). Finally, the drop at its last point (already in 2026) should be viewed with caution. In the case of Google Trends series, drops to the far right occur due to the same reasons, such as an incomplete time window (period still consolidating) or recent variations in interest.

Therefore, methodologically, it is good to leave the final segment "provisional" and record in the text that it is the exact date of collection and the selected interval that determines the reading of the end of the series (GOOGLE TRENDS, 2026).

According to the study's proposal, these findings support two conclusions regarding topic 4.3: (i) the international leadership of nations and entities in the scientific production process (evaluated via bibliometrics) occurs simultaneously with a dynamic of global public interest that can expand with "waves" and "jumps" related to events and changes in adoption, and (ii) XAI solidifies as a topic not only studied in universities but also demanded in society, which explains its transition from a "desirable" requirement to a "necessary" requirement in high-impact applications. Complementing a reading of interest over time represented earlier in Figure 2, the second piece of evidence we extracted from Google Trends tells us which terms people actually search for when "explainable artificial intelligence" is the topic, separating more frequent queries (of stable volume) from queries on the rise (recent growth).

ISSN: 2675-6218 - RECIMA21
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

This layer is vital to the objective of our article, as it allows insights into why interest increases or decreases and questions about the multiple meanings of the term "XAI" in the public domain — which, in turn, informs our analysis of the "distribution of queries about XAI in the world" (GOOGLE TRENDS, 2026).

Considering the typical queries in the current research, it is possible to infer that the most frequently asked terms here are highly broad and conceptual: "AI", "explainable", "explainable AI", "explainability", and so on. This pattern suggests that the vast majority of the audience may be being introduced to the topic in its generic (AI) or exploratory (explainability) variety in their daily experience, and that both types of data are aligned with the discussion about transparency and interpretability when applied in continuous use (Sharma *et al.,* 2024).

Although this is where "XAI" actually appears or doesn't appear in these counts, its relative decrease in the list is significantly different from the high variance of "explainability," as are types with significant growth similar to combinations like "explainable AI" and "AI explainability," which also show a significant increase (e.g., +200%−+250% in capture).

Interpretively, this means that the public wants more abstraction or elucidation than an acronym for XAI, and this becomes more evident when the acronym begins to be confused with definitions beyond some academic ones (GOOGLE TRENDS, 2026). The range of terms is even clearer when looking at the ascending query column; terms like "OpenAI", "XAI Musk", "Elon Musk XAI", "XAI crypto ", "XAI token", "XAI coin ", " Grok ", and "XAI website" are heavily sprinkled with "Great Increase". This sequence of blocks proves that the search spikes for 'XAI' in the recent period are not so much in XAI itself (the new word) as in the xAI brand/company and related topics (product, news, cryptocurrency assets), especially as it relates to the academic term XAI as Explainable Artificial Intelligence. Translation: The term "XAI" is semantically ambiguous and lacks context when used in the public domain, which can lead to inflated or distorted measurements of interest if the collection only describes XAI in the public domain (GOOGLE TRENDS, 2026).

This observation may explain — and may be related to — why interest can be shown as jumps on a timeline graph, due to the fact that the peaks are not so much a result of interest in "explainability" as a result of events or media attention to " xAI " (brand).

From a research perspective, we could also conclude that, for the article, it is more appropriate to carefully read the "query distribution," where Google Trends can indicate general social demand and scientific leadership according to the country, where scientific influence should probably be presumed based on bibliometric information (publications, citations, networks) reported from a specific XAI analysis (Sharma *et al.,* 2024).

Methodologically, it also demonstrates that the findings can be further refined: in areas where the goal is to spark interest in Explainable AI (scientific field), we would suggest classifying less ambiguous expressions (e.g., explainable AI, explainability, or explainable machine learning)

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

and/or using combinations/segments to minimize noise (e.g., comparing "explainable AI" with "XAI" and reporting different results).

Therefore, the study presents consistency between the indicator's purpose (showing interest in academic XAI) and the true measured value (academic interest vs. convergent brand/event interest) (GOOGLE TRENDS, 2026). Thus, section 4.3 brings together both pieces of evidence, namely a timeline and related queries, concluding that XAI is still a common research topic, but the increase in recent years can be attributed to the competition associated with the definitions of XAI. This does not dilute the signal; on the contrary, it broadens its reach: it presents the explainability agenda as growing along with the "popularization" of AI vocabulary—therefore, being cautious regarding terms across countries, trends, and eras (Sharma *et al.,* 2024; GOOGLE TRENDS, 2026).

CONSIDERATIONS

Our findings indicate that the transition from traditional AI to XAI is being driven not only by advanced algorithmic innovation, but by the intersection of multiple factors: first, more complex and opaque models; second, more concentrated high-impact applications; third, concerns about bias and algorithmic fairness; and fourth, regulatory and institutional pressures for transparency and accountability.

Bibliometric analysis demonstrates the emergence of XAI as a scientific agenda, particularly the acceleration of a body of publications since 2018 in the space and the prominent position that specific authors and editorial sources occupy in this field. Furthermore, Google Trends results suggest a growing social demand, but also point to an important methodological issue: when searching for information, the term "XAI" can be used for non-academic meanings, so the use of misleading terminology (such as "explainable AI" and "explainability") should be avoided in the hope of quantifying public interest in the scientific concept in question.

Bibliometric conclusions depend on the database, time period, and search strategy, and are therefore subject to limitations, while search indicators suggest relative interest and may therefore be subject to semantic noise. For the future agenda, we propose expanding our bibliometric analysis with co-authorship and co-citation networks and comparative trends by applied fields (healthcare, finance, public sector) which would further improve the integration of adoption metrics/standards/guidelines/tools, potentially providing a better map of how explainability is shifting from research to AI design and governance.

ISSN: 2675-6218 - RECIMA21
Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

REFERENCES

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**, v. 104, n. 3, p. 671–732, 2016.

BISHOP, Christopher M. **Pattern Recognition and Machine Learning**. New York: Springer, 2006.

DOSHI-VELEZ, Finale; KIM, Been. Towards a rigorous science of interpretable machine learning. **arXiv preprint**, 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. **Explainable artificial intelligence (topic)**. Scope: Global; time frame: last five years; type: Google Web Search. Accessed: February 18, 2026.

GUNNING, David. **Explainable Artificial Intelligence (XAI)**. Arlington: DARPA, 2017.

LECUN, Yann; BENGIO, Yoshua; HINTON, Geoffrey. Deep learning. **Nature**, 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**. 2nd ed. [S. l.]: Leanpub, 2021.

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. 3rd ed. Upper Saddle River: Pearson, 2016.

SHARMA, Chetan; SHARMA, Shamneesh; SHARMA, Komal; SETHI, Ganesh Kumar; CHEN, Hsin-Yuan. Exploring explainable AI: a bibliometric analysis. **Discover Applied Sciences**, v. 6, 2024.

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.

REVISTA CIENTÍFICA - RECIMA21 ISSN 2675-6218

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**, v. 7, n. 2, p. 76–99, 2017.

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

ISSN: 2675-6218 - RECIMA21

Este artigo é publicado em acesso aberto (Open Access) sob a licença Creative Commons Atribuição 4.0 Internacional (CC-BY), que permite uso, distribuição e reprodução irrestritos em qualquer meio, desde que o autor original e a fonte sejam creditados.