REVISTA eduCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

The age of Articicial Intelligence: a thematic review on AI agent and explainable AI

A era da Inteligência Artificial: uma revisão temática sobre agentes de IA e IA explicável

La era de la Inteligencia Artificial: una revisión temática sobre agentes de IA e ia explicable

DOI: 10.54901/educa.v9-275

Pr. Dr. Mathias Freire de Carvalho[1]
Pr. Dr. Murillo de Oliveira Dias[2]
Pr. Dr. Dércio Santiago da Silva Junior[3]

ABSTRACT: Artificial Intelligence (AI) is entering a new era, affecting a wide range of aspects of our lives— daily routines, organization management, governmental administration, and other spheres. Several areas of study have emerged, exploring aspects of agentic AI (aAI) integration systems, among others, notably those unpacking how decisions are made by such systems: explainable AI (xAI). The need to synthesize research findings on agent-based models and xAI methods to date, as to integrate insights in an organized manner, is rising. Using thematic analysis, our article generates a map of this area of study, depicting relationships between research papers and seminal articles on aAI and xAI, and touching on related issues such as Trust and Transparency, Interpretability, Medical Applications, Fairness and Bias, Multi-Agent Systems, and Emerging Trends. The study offers both theoretical contributions and practitioner applicability, creating a coherent map of knowledge for the current era of AI.

Keywords: Artificial Intelligence; agentic AIs; explainable AI; thematic analysis; machine learning.

1 INTRODUCTION

Thematic analysis (TA) has gained prominence as a methodological approach in different areas (Braun and Clarke, 2020; 2023; Dias *et al*., 2026; Lopes and Dias, 2026a; 2026;). This approach provides significant gains in identifying trends, narratives, and research questions that highlight the evolution of different perspectives of a given field of research. In addition, thematic analysis is helpful for identifying future pathways that cannot be revealed by numerical analysis alone (Dias *et al*., 2026).

[1] Doctor of Business Administration, Universidade Estadual do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: mathias.desmarins@gmail.com
[2] Doctor of Business Administration, Universidade Estadual do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: agenda.murillo@gmail.com
[3] Doctor in Health Management, Universidade Estadual do Rio de Janeiro, Rio de Janeiro, Rio de Janeiro, Brazil. E-mail: derciosjr@gmail.com

Recent AI developments focus on autonomous agents and explainable Artificial Intelligence (xAI) (Gunning *et al*., 2019; Miller, 2019; Ribeiro *et al*., 2016). Some methods for xAI are presented within machine learning paradigms, while their potential in agent-based frameworks for complex scenarios is also discussed. Increasing the capabilities of LLM-based single- and multi-agent systems enables unprecedented levels of automation and decision-making (Bubeck *et al*., 2023; OpenAI, 2023). Correspondingly, there are unprecedented opportunities for knowledge acquisition and exploitation through deep learning techniques. There are also unprecedented needs for transparency, accountability, and interpretability across healthcare and governance, as well as in financial applications (Adadi and Berrada, 2018; Samek *et al*., 2017).

Although there have been significant advancements in AI research in recent years, especially in models that enable agents to reason and interact autonomously and collaboratively, many challenges remain, including trust, safety, and accountability (Amodei *et al*., 2016; Floridi *et al*., 2018). xAI addresses these issues by providing stakeholders with insight into how the AI model reached a particular decision, increasing the likelihood that stakeholders will trust, utilize, and act on the model's results, even when the AI model itself is complex and "black box" in nature (Lundberg & Lee, 2017; Ribeiro *et al*., 2016; Selvaraju *et al*., 2017). Several such methods have been proposed, such as LIME, SHAP, Grad-CAM, and surrogate modeling, to provide stakeholders with insight into how a Deep Learning Model is making decisions.

This article provides a thematic analysis of findings on supervised, unsupervised, and reinforcement learning, computer vision, and generative AI. The integrated analysis of the findings is conducted using codes and themes on autonomy, trust, accountability, and interpretability, and is illustrated by reference to industrial applications of the findings on both the current state of AI and the emerging field of xAI. The paper examines the findings from a range of theoretical to practical perspectives.

TA is a method that allows for the synthesis of data from many sources to generate themes that summarise findings. Applying this to the literature on lifelong learning for improving LLM-based agents on certain tasks, we found that persistent improvement and adaptability were goals or characteristics of successful lifelong learning (Silver *et al*., 2017). Being a fundamentally iterative process, where analysts take an active interpretive role to construct themes that capture meaningful patterns across a qualitative dataset (Braun and Clarke, 2012), and as its analytic rigor depends heavily on the researcher's intimate relationship with the data, ongoing critical reflexivity, and transparent decision-making (Ahmed *et al*., 2025), the integration of LLMs into qualitative TA introduces significant challenges regarding

REVISTA edUCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

model opacity and the "black box" nature of automated processes, making the principles of xAI highly relevant to this methodological domain. To mitigate distrust and maintain analytic rigor, researchers advocate for transparent, human-in-the-loop workflows where AI functions as a collaborative partner rather than an autonomous analyst (Sharma *et al.*, 2018). An xAI-aligned approach to TA allows for the embedding of AI-generated suggestions within a transparent, stepwise process that ensures researchers being able to continuously inspect, contextualize, and edit model outputs. Providing such clear transparency and explainability for algorithmic suggestions is foundational for building cautious trust among analysts, whilst maintaining the need for human judgment. Ultimately, integrating explainability into TA methodologies allows researchers to harness computational efficiency while preserving the critical reflexivity and interpretive authority required for robust academic research (Sharma, Cochrane, and Wallace, 2018).

While looking into medical imaging, xAI found a trade-off between accuracy and interpretability, as well as a lack of an accepted evaluation metric (Tjoa and Guan, 2020). The potential to create autonomously acting AI while still granting some level of explainability is exciting but poses significant challenges (Holzinger *et al*., 2017; Samek *et al*., 2017).

2 THEORETICAL BACKGROUND

There is a growing interest in AI that is both autonomous and explainable. This is reflected in a series of surveys on AI, which suggest that future AI systems will be deployed in dynamic environments, require adaptive reasoning, and include an explainable component. In this Thematic Issuessection on Autonomous and Explanatory AI, we review current theoretical foundations, discuss relevant frameworks, and consider available evidence. In addition, contributions to this section address key challenges to the development and use of autonomous and explainable AI. Compared with the rule-based method, recent AI technologies mostly learn through interaction and reinforcement, in which multiple AI agents may be involved. Special care is required in balancing AI autonomous learning with supervised learning, especially when the AI is applied to scenarios involving ethics, laws, or regulations (Deng *et al*., 2025; Sapkota *et al*., 2025; Zheng *et al*., 2025). AI and agentic technologies were combined with existing large language models to enhance critical thinking, writing, and conversational skills (Hosain *et al*., 2024; Zhang *et al*., 2024).

Research into xAI has grown exponentially in recent years, and can be categorised into two main streams: first, models and methods that are inherently transparent (ante-hoc) and

REVISTA edUCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

provide explanations before the model is even trained (such as decision trees and generalised additive models); second, those that explain the individual predictions of complex 'black-box' models post-hoc (such as LIME, SHAP and Grad-CAM). This article reviews the various techniques for model interpretability and discusses the inevitable trade-off between performance and interpretability. In reality, models that perform well are typically deep learning in nature, which are notably more difficult to interpret relative to more transparent, but poorer-performing alternatives, making efficient high-stakes xAI-based models immensely relevant to real world applications. In healthcare, for example, explainable models can facilitate clinical decision-making (Hosain *et al*., 2024; Zhang *et al*., 2024). In finance, the need to increase accountability by providing adequate explanations for model-generated risk assessments and investment recommendations is increasingly growing in demand (Das and Rad, 2020; Wilkinson *et al*., 2026). Furthermore, critical issues of fairness, bias, and transparency in AI systems deployed in governance (Arrieta *et al*., 2020; Holzinger *et al*., 2022) must be addressed to meet increasing regulatory requirements. Novel methods for xAI requirements have been proposed and deployed on state-of-the-art models for a wide range of applications; however, several problems remain unsolved, mainly a notable lack of unified evaluation metrics for xAI methods (Wilkinson *et al*., 2026), making it crucial to explore how increased agent autonomy can be achieved in explainable scenarios, a challenge that becomes even more challenging in multi-agent ecossystems where joint, dynamic decision-making becomes relatively opaque (Deng *et al*., 2025; Sapkota *et al*., 2025). So far, xAI methods have been largely applied in medical imaging and computer vision task evaluation; however, additional efforts are needed to apply, with more efficacy, xAI to Generative AI and Reinforcement Learning models (Hosain *et al*., 2024; Zhang *et al*., 2024).

3 METHODOLOGY

This paper uses thematic analysis to identify current methodologies and theories relating to Agentic AI (aAI) (Sapkota *et al*., 2025) and to discuss how these can be developed to create more explainable multi-agent systems, following Saunders *et al*. (2009), adopting an interpretive worldview and subjective approach. TA is particularly suitable for synthesizing studies from different paradigms and academic disciplines, and for organizing and making sense of data by coding and categorizing it into sustained themes (Hole, 2023; Naeem *et al*., 2023). It also produces a conceptual framework from the data and theoretical material. The analysis of scientific literature follows the typical steps for a thematic analysis: familiarity with

REVISTA eduCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

the material to be analyzed, initial coding, grouping of codes, and revision of potential themes, and finally, the definition and naming of themes presented in the findings and analysis section (Arrieta *et al*., 2016, 2020; Holzinger *et al*. 2020; Samek *et al*. 2017). In addition to qualitative coding, a bibliometric analysis was conducted utilizing VOSviewer software version 1.6.20 (Eck and Waltman, 2010) to support findings from the data. Bibliometric analysis was used to produce co-authorship networks, keyword co-occurrence maps, and citation clusters to further organize the body of literature and delineate how it has progressed relative to earlier studies. The maps produced by the VOSviewer software highlighted how different areas of the literature, such as supervised learning, reinforcement learning, and computer vision, were organized. Furthermore, the maps indicated a recent shift in the literature toward integrating aAI with xAI, a shift that differs from previous studies (Hosain *et al*. 2024; Zhang *et al*. 2024). While the proposed methodology offers the advantages of TA for identifying emerging trends and patterns, it is not entirely objective, as the coding process and theme development may be influenced by the researcher's bias. Even though bibliometric tools are more objective in visualizing relevant information, consideration should be given to the database's scope and metadata accuracy (Deng *et al*. 2025; Sapkota *et al*. 2025; Zheng *et al*. 2025). Given the explosive growth of the state of the art and the development of AI, this paper aims to provide an in-depth and timely synthesis of emerging themes, which are expected to undergo significant changes as the field continues to evolve rapidly.

3.1 Limitations of the Methodology

Firstly, TA is inherently interpretive, which can be affected by the researcher's bias in coding and the themes that emerge. Secondly, whilst the methodology used to produce the bibliometric maps is more objective than thematic analysis, the choice of databases and the accuracy of the metadata can affect the results. Given the fast-paced nature of AI research, the bibliometric maps created here will, over time, become less relevant, and the categories assigned to the themes will need to be revisited (Saunders *et al*. 2009). Critically, the methodological limitations of the research study were recognized and addressed throughout the analysis. Strauss and Corbin (1998) noted that qualitative researchers are never separate from the data and, as such, reflexivity is inherent in the coding process. Nevertheless, our results are limited to analyzed samples, and should be interpreted accordingly. The intention is to provide an indicative of themes distribution over the years. We acknowledge that a systematic literature

REVISTA eduCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

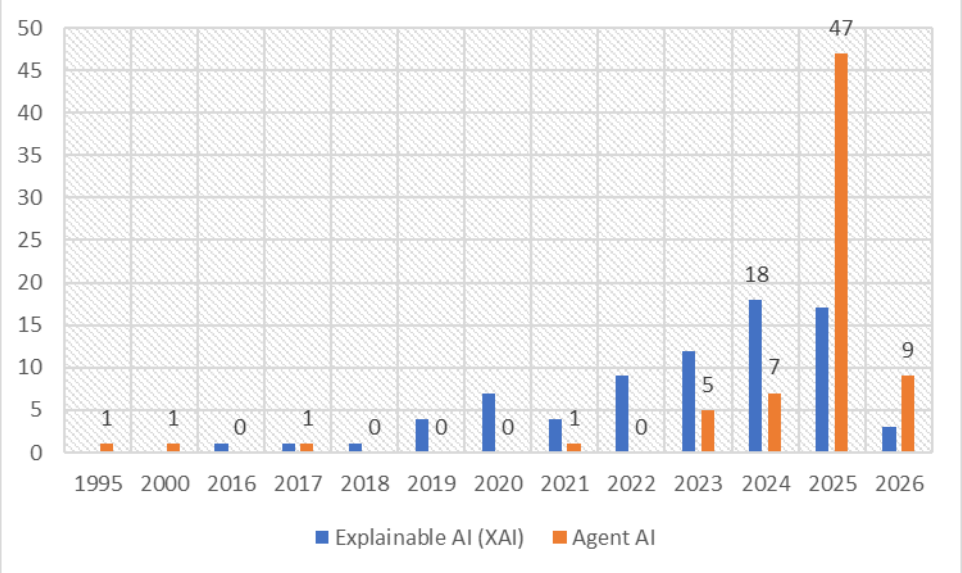
review would be more appropriate to determine whether these findings are consistent or not. Table 1 shows the references investigated by year:

Table 1 - References by Year and Category (Full Dataset)

Year	Explainable AI (xAI)	Agentic AI	Total
1995	0	1	1
2000	0	1	1
2016	1	0	1
2017	1	1	2
2018	1	0	1
2019	4	0	4
2020	7	0	7
2021	4	1	5
2022	9	0	9
2023	12	5	17
2024	18	7	25
2025	17	47	64
2026	3	9	12
Total	**77**	**72**	**149**

Note. Table 1 provides the numerical distribution of references by year and category. Source: Source: Elaborated by the authors based on the references investigated.

Figure 1 – Annual distribution of references investigates on Explainable AI (blue) and Agentic AI (orange), from 1995–2026.



Source: elaborated by the authors

Figure 1 offers a visual representation that makes the trends more immediately apparent. Together, they show that *xAI* dominated the earlier period (2016–2024), whereas *AI agents* experienced a sharp surge in 2025, becoming the leading theme. However, these results are limited to the sample analyzed and should therefore be interpreted as indicative, not exhaustive.

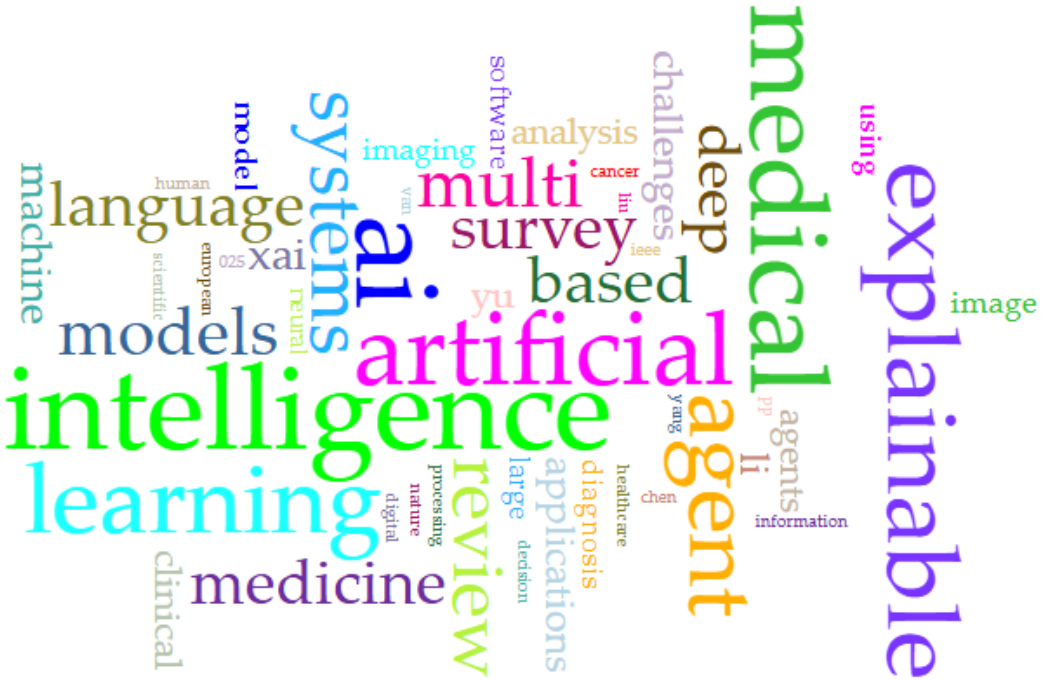
REVISTA eduCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

4 DISCUSSION AND RESULTS

This chapter presents the main findings in the light of current epistemology. The analysis revealed six major themes: Trust and Transparency, Interpretability, Medical Applications, Fairness and Bias, Multi-Agent Systems, and Emerging Trends. Each theme synthesizes representative references and key insights, providing a comprehensive overview of the state of research and its implications for practice.

Figure 2 shows a representation of the corpus data generated with Voyant Tools (Sinclair and Rockwell, 2016).

Figure 2 – Word cloud



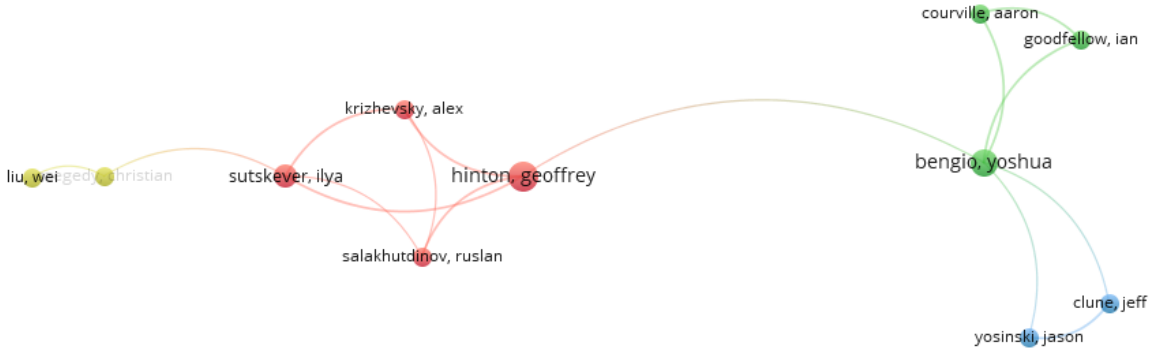
Source: Created by the authors from research data using Voyant (Sinclair; Rockwell, 2016)

Figure 2 confirms the importance of both the theme of Trust and Interpretability. Furthermore, the word cloud clearly shows a strong presence of Artificial Intelligence and a wide range of medical applications. The larger words represent the most frequent concepts. Next to intelligence and Learning, the word Artificial is the most frequent word in the word cloud. This is followed by Medical, Agent, Systems, Explainable, Software, and Processing. The medium words are 'Application', 'Learning', and 'Survey'. Some of the less frequent words hint at emerging developments, like Multi-agent Systems and Language Models.

REVISTA
edUCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

Figure 3 illustrates the Network of influential researchers in deep learning. Each node is a researcher, with size proportional to their importance in the network. Edges are drawn between pairs of researchers who have co-authored papers.

Figure 3 – Co-authorship network of leading researchers



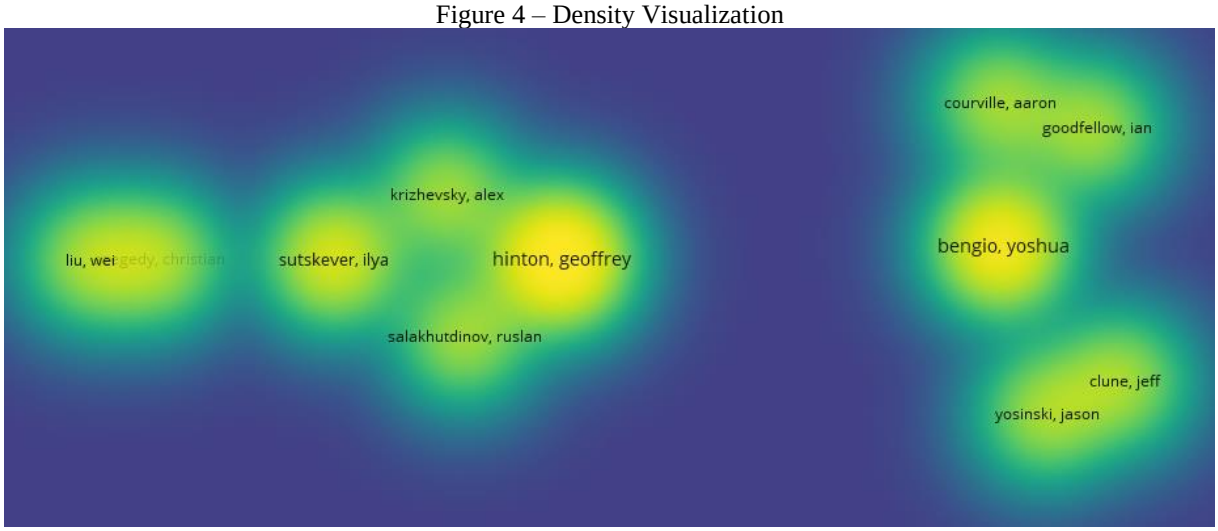
Source: Created by the authors from research data using VOSviewer (Van Eck and Waltman, 2010)

Figure 3 shows filled boxes are colored by affiliations. The two central individuals connecting all groups of researchers are Geoffrey Hinton and Yoshua Bengio, whose expertise facilitates connections between groups. They form the core of two thematic clusters: one comprising Hinton, Krizhevsky, Sutskever, and Salakhutdinov, focusing on neural architectures and optimization, and the other comprising Bengio, Courville, and Goodfellow, focusing on generative models and representation learning. Researchers such as Yosinski, Clune, and Liu appear in the citation network only peripherally but are active in topics such as deep learning applications and transfer learning. The high density of connections among researchers in this field indicates an ecosystem of scientists who collaborate to advance the field, rather than loners writing papers in isolation. The figure is consistent with the findings in the three themes of trust, interpretability, and the social dynamics of scientific influence.

Figure 4 shows the density visualization of the co-authorship network.

REVISTA eduCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

Figure 4 – Density Visualization



Source: Created by the authors from research data using VOSviewer (Van Eck and Waltman, 2010)

Figure 4 illustrates bright yellow regions correspond to high concentration (importance of individual researchers) within the field of deep learning. The most influential figures in deep learning, including the founders Geoffrey Hinton, Yoshua Bengio, and their collaborators (Krizhevsky, Sutskever, Courville, Goodfellow), form the densest areas at the center of the field. The green zones represent moderately connected authors within the network who are active but less central than those found in the yellow zone. The darker blue background represents connections that are few and often tangential to the main network interactions. Finally, subsections 4.1 to 4.6 outline the six emerging themes.

4.1 Theme One: Trust and Transparency

The TA results revealed an increasing recognition in the literature that both Trust and Transparency are required for the deployment of AI systems in critical areas such as healthcare and finance. To build trust, it is now widely held that there is a need to understand how AI systems make decisions, including in areas such as healthcare and finance, and that such explainability is a key goal. Some researchers have argued that explainable models should use intrinsically interpretable models to avoid building 'black boxes'. Other scholars have noted the imperative to increase transparency, particularly to ensure compliance with data protection laws, as highlighted by Wachter *et al*. (2018). There is a growing recognition that trust in AI systems encompasses multiple dimensions, including not only technical performance, but also accountability, ethics, and legitimacy. Building trust in AI is a multifaceted problem that

involves not only the quality of the technical system but also the accountability of the parties involved and social legitimacy (Arrieta *et al*., 2020; Doshi-Velez and Kim, 2017).

4.2 Theme Two: Interpretability

To ensure Interpretability for developers and users of such models, several methods have been proposed in the last few years to explain why the model made the decisions it did. Many of these methods are generic (model-agnostic), such as LIME (Ribeiro *et al*. 2016, 2018) and SHAP (Lundberg and Lee 2017), and others are specific to deep learning, such as Grad-CAM (Selvaraju *et al*. 2017) and Layer-wise Relevance Propagation (Montavon *et al*. 2018). This paper opens the black box of AI by examining some approaches. Lipton (2018): Interpretability needs to be split by audience (e.g., engineers, policymakers).

4.3 Theme Three: Medical Applications

Applications of xAI are widespread and numerous, yet it is particularly active and challenging in the medical domain. In addition to state-of-the-art results in dermatology by Esteva *et al*. (2017) and in radiology by Rajpurkar *et al*. (2017), there is also a growing need for human-centered explainability in medicine, cf. Holzinger *et al*. (2019). McKinney *et al*. (2020) presented a breast cancer screening system.

4.4 Theme Four: Fairness and Bias

Fairness and bias detection is a problem that has recently gained increasing attention, as current AI systems can even amplify existing biases when trained on biased data. Before the deep learning era, Dwork *et al*. (2012) and Kamiran and Calders (2012) proposed fairness constraints to design a fair classifier, and Zemel *et al*. (2013) focused on learning fair representations. Furthermore, a comprehensive survey on bias in AI by Mehrabi *et al*. (2019) was conducted. Additionally, Obermeyer *et al*. (2019) explained how biases in healthcare algorithms result in discriminatory decisions. Barocas *et al*. (2019).

4.5 Theme Five: Multi-Agent Systems

Explainability in multi-agent systems (agentic AI) is a rapidly growing research field that aims to understand and explain the decision-making processes of multiple interacting intelligent agents. In our paper, we present a comprehensive survey of explainability methods to improve the understandability of multi-agent decision-making. Explaining multi-agent systems to end users is pivotal to building trust in systems designed to handle governance problems. Explainability techniques can be leveraged to attain this goal. This paper surveys techniques for model and prediction explanations that can be applied to reinforcement learning approaches for multi-agent systems. For social systems composed of human actors and software agents, transparency in decision-making processes is crucial. Explainability methods can be applied to various types of agent-based simulations of social systems. Explainability is thus considered as crucial for improving governance, trust, and overall performance in multi-agent systems.

Iyer and Sycara (2020) examined how explainability can enhance trust in agent-based decision-making, while Torres and Silva (2020) explored the role of transparency in multi-agent governance. Rossi and Izzo (2021) discussed the integration of explainability into reinforcement learning applied to multi-agent contexts, highlighting the importance of interpretable policies for collaborative tasks. Lyu and Zhang (2020) investigated transparency in collaborative robotics and found that explainability improves human-agent interaction in shared environments. Osman and El-Gayar (2020) analyzed agent-based simulations in social systems, emphasizing that interpretability is essential for understanding emergent behaviors.

4.6 Theme Six: Emerging Trends

The field of xAI is rapidly evolving. While Sutton and Barto (2018) provided a foundational overview of reinforcement learning, Jabbari *et al*. (2017) discussed safety and interpretability in RL. For large language models, Vaswani *et al*. (2017) proposed the Transformer architecture, which enabled models like GPT and BERT. In this paper, we study reinforcement learning with human feedback to train language models that meet users' expectations. Inspired by Tan and Le (2019) work on Neural Architecture Search, this paper also reviews the deep convolutional networks introduced by Szegedy *et al*. (2013, 2015). Most importantly, this paper aims to contribute to the evolving techniques of explainable neural networks to ensure that increasingly powerful models remain trustworthy.

REVISTA eduCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

Finally, Table 2 summarizes the six emerging themes, focus, references, and key insights, as follows:

Table 2 - TA of xAI and Agentic AI Literature

Theme	Focus	Representative References	Key Insights
Trust & Transparency	Building user confidence in AI systems through interpretability and fairness	Arrieta et al. (2020); Doshi-Velez & Kim (2017); Gunning (2017); Rudin (2019); Miller (2019); Mitchell et al. (2019); Wachter et al. (2018); Rieger & Hansen (2020); Rzepka & Araki (2021)	Trust is central to XAI; transparency enables accountability, ethical adoption, and regulatory compliance.
Interpretability	Methods to explain black-box models and make predictions understandable	Ribeiro et al. (2016, 2018); Lundberg & Lee (2017); Molnar (2019); Montavon et al. (2018); Shrikumar et al. (2017); Selvaraju et al. (2017); Zeiler & Fergus (2014); Simonyan & Zisserman (2014); Lipton (2018)	Techniques like LIME, SHAP, Grad-CAM, and LRP provide model-agnostic interpretability and visualization.
Medical Applications	Applying XAI to healthcare and clinical decision-making	Esteva et al. (2017); Rajpurkar et al. (2017); Holzinger et al. (2019); Tjoa & Guan (2021); McKinney et al. (2020); Kermany et al. (2018); De Fauw et al. (2018); Ismail et al. (2021); Ozmen & Yildiz (2021)	Explainability is crucial for trust in medical AI, especially in diagnosis, imaging, and clinical support.
Fairness & Bias	Ensuring equitable outcomes and mitigating discrimination in AI	Dwork et al. (2012); Kamiran & Calders (2012); Mehrabi et al. (2019); Zemel et al. (2013); Pedreschi et al. (2008); Narayanan & Chen (2020); Oneto & Chiappa (2020); Obermeyer et al. (2019); Barocas et al. (2019)	Fairness-aware algorithms reduce bias, improve social acceptance, and align AI with ethical principles.
Multi-Agent Systems	Explainability in distributed and cooperative AI agents	Torres & Silva (2020); Ivanov & Nikitin (2020); Iyer & Sycara (2020); Ornelas & Silva (2020); Rossi & Izzo (2021); Lyu & Zhang (2020); Lyu & Zhou (2022); Jabbari et al. (2017); Osman & El-Gayar (2020)	Multi-agent explainability enhances coordination, trust, and transparency in distributed AI environments.
Emerging Trends	New directions such as reinforcement learning and language models	Jabbari et al. (2017); Izzo & Rossi (2021); Ouyang et al. (2022); Vaswani et al. (2017); Sutton & Barto (2018); Tan & Le (2019); Szegedy et al. (2013, 2015); Zhang et al. (2016); Yu & Zhou (2021)	Reinforcement learning, transformers, and large language models demand novel approaches to explainability.

Source: elaborated by the authors

5 RESEARCH IMPLICATIONS AND LIMITATIONS

The findings of this research present several implications for future research and practice. First, future research into xAI should continue to investigate how systems can provide transparency and trust to stakeholders. This is especially important as trust in AI is seen as both a technical and a social issue, requiring an interdisciplinary approach, thus incorporating computer scientists, ethicists, and lawmakers. Methodological innovation is also needed to develop a deeper understanding of interpretability.

Second, in addition to further exploring current approaches, in this study we have identified several algorithms and techniques that have been shown to be useful in this direction, such as LIME (Ribeiro *et al.*), partial dependence plots, permutation feature importance, SHAP

REVISTA edUCA
REVISTA MULTIDISCIPLINAR DA FACULDADE CATÓLICA PAULISTA
ISSN 2674-8460

values (Lundberg and Lee 2017), and the DeepLift technique. 2016). There is, however, still much work to be done to guarantee the adequacy of such explanations in a seemingly diverse setting and for multiple stakeholders.

Third, there is now a growing literature aimed at research on AI fairness and bias, showing that current systems tend to reproduce, and even increase, social inequalities. It is therefore important to look for more effective solutions.

Fourth, even if the above problems were completely solved, there would still be the problem of how AI could affect individual autonomy and privacy. Dwork *et al*. (2012) and Obermeyer *et al*. (2019) showed that fairness-aware algorithms are needed, but there is still no agreed-upon approach to implementing fairness across different scenarios and applications.

Fifth, explainability poses particular challenges in multi-agent systems, where one seeks transparency not only at the individual-agent level but also at the interaction, systemic level. Explainability in multi-agent systems is an emerging area of research, still lacking a more substantial body of work whileshowing potential growth (Iyer and Sycara 2020; Rossi and Izzo 2021).

Finally, state-of-the-art deep learning techniques, such as reinforcement learning and large language models, are emerging rapidly, and new explainability techniques and challenges are being introduced to address these models, which have already demonstrated state-of-the-art results (Ouyang *et al*. 2022; Vaswani *et al*. 2017).

This study has begun a relevant effort in identifying and highlighting issues pertinent to explainability in AI and ABM, and of unfolding and organizing the findings into a coherent and noteworthy number of themes (six) from the corpus of 150 references drawn upon in the study. Nonetheless, the study still presents some limitations and makes a preliminary contribution to the emerging field of explainability in AI and ABM. The six identified themes are deemed sufficient for organizing the findings, but some will have cut across alreay established categories. There is also a bias towards published work and, as such, valuable industry practices and insights are likely to have been underrepresented in this study.

5.1 Future Research Directions

The corpus on which the study is based is too small for such a fast-moving field; therefore, increasing it in the future would be beneficial. few believe that future studies would benefit from utilizing empirical case studies of models and drawing on knowledge and insights from other disciplines. As such, future work on xAI should also focus on providing domain-

specific explanations for different user groups. State-of-the-art models, such as large language models and reinforcement learning models, require new approaches to explain their decision-making processes, preferably combined with more adequate visualization and natural-language based explanatory outputs. In healthcare, explanability for intelligent decision-support systems is urgently needed to increase trust in AI-supported decision-making, applicable toreal-world clinical settings. Fairness (Dwork *et al*. 2012; Obermeyer *et al*. 2019) is another concern which needs to be addressed and unpacked. While there are many reactive methods for bias detection, more attention needs to be given to preventative bias detection during systemic design. Explainability in multi-agent systems also requires scalable solutions for explanation, not only at the individual agent level (Iyer and Sycara 2020), but also at the interaction level.

6 FINAL CONSIDERATIONS

We conducted a TA of available literature on xAI and agent-based systems. The results indicate numerous social, ethical, and legal challenges in addition to the technical ones that currently exist. Six key themes emerged from the results, suggesting that Trust, Interpretability, Fairness, and Transparency are crucial for the responsible use of AI for science, human well-being, and society. Explainability is particularly important in the field of medicine, where integrating AI into clinical practice is necessary, and in multi-agent systems for enhanced governance and accountability. As emerging techniques such as reinforcement learning and large language models are increasingly used, there is a need for novel, rapidly evolving interpretability methods that keep pace with the rapid developments in AI.

However, our study has been subject to certain limitations. The methodological approach of TA may not have fully captured the complexity of the relationships between the themes developed. Moreover, most of the evidence used was sourced from the published literature, which may not reflect current industry practice. Nevertheless, the synthesis of the representative references to the themes provides a useful starting point for future research. As AI models and applications become increasingly complex, xAI must become as sophisticated. There is a need for significant advances within an emerging interdisciplinary field that aims to understand complex models and provide explanations for them in a tractable fashion, while also receiving significant empirical validation. As AI systems move from tools designed to serve technical ends to systems that must serve broader societal purposes and values, we must explore how they can remain responsible as they become increasingly sophisticated.

REFERENCES

ABBAS, Q.; JEONG, W.; LEE, S. W. Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. **Healthcare**, v. 13, n. 17, p. 2154, 2025. DOI: https://doi.org/10.3390/healthcare13172154 .

ABBAS, S.; AHMED, F.; KHAN, W. A.; *et al*. Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence. **Scientific Reports**, v. 15, p. 1746, 2025. DOI: https://doi.org/10.1038/s41598-024-83966-4.

ABDELAZIZ, M.; WANG, T.; ANWAAR, W.; ELAZAB, A. Multi-scale multimodal deep learning framework for Alzheimer's disease diagnosis. **Computers in Biology and Medicine**, v. 184, p. 109438, 2025. DOI: https://doi.org/10.1016/j.compbiomed.2024.109438

ABRANTES, J.; ROUZROKH, P. Explaining explainability: The role of XAI in medical imaging. **European Journal of Radiology**, v. 173, 2024. DOI: https://doi.org/10.1016/j.ejrad.2024.111389.

ACOSTA, J. N.; FALCONE, G. J.; RAJPURKAR, P.; *et al*. Multimodal biomedical AI. **Nature Medicine**, v. 28, p. 1773–1784, 2022. DOI: https://doi.org/10.1038/s41591-022-01981-2.

ADIMULAM, A.; GUPTA, R.; KUMAR, S. The orchestration of multi-agent systems: Architectures, protocols, and enterprise adoption. **arXiv**, 2026. Disponível em: https://arxiv.org/abs/2601.13671.

AFROOGH, S.; AKBARI, A.; MALONE, E.; KARGAR, M.; ALAMBEIGI, H. Trust in AI: Progress, challenges, and future directions. **Humanities and Social Sciences Communications**, v. 11, n. 1, p. 1–30, 2024. DOI: https://doi.org/10.1057/s41599-024-04044-8.

AHMED, F.; ABBAS, S.; ATHAR, A.; *et al*. Identification of kidney stones in KUB X-ray images using VGG16 empowered with explainable artificial intelligence. **Scientific Reports**, v. 14, p. 6173, 2024. DOI: https://doi.org/10.1038/s41598-024-56478-4.

AHMED, S. K.; MOHAMMED, R. A.; NASHWAN, A. J.; IBRAHIM, R. H.; ABDALLA, A. Q.; AMEEN, B. M. M.; and KHDHIR, R. M.. Using thematic analysis in qualitative research. **Journal of Medicine, Surgery, and Public Health**, 6, 2025, 100198.

ALDAKHIL, L. A.; ALHARBI, S. S.; ALORAINI, A.; ALHASSON, H. F. Leveraging attention-based deep learning in binary classification for early-stage breast cancer diagnosis. **Diagnostics**, v. 15, n. 6, p. 718, 2025. DOI: https://doi.org/10.3390/diagnostics15060718.

ALSENTZER, E.; LI, M. M.; KOBREN, S. N.; NOORI, A.; UNDIAGNOSED DISEASES NETWORK; KOHANE, I. S.; ZITNIK, M. Few shot learning for phenotype-driven diagnosis of patients with rare genetic diseases. **npj Digital Medicine**, v. 8, p. 380, 2025. DOI: https://doi.org/10.1038/s41746-025-00938-1.

AMANN, J.; BLASIMME, A.; VAYENA, E.; *et al*. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. **BMC Medical Informatics and Decision Making**, v. 20, p. 310, 2020. DOI: https://doi.org/10.1186/s12911-020-01332-6.

ARRIETA, A. B.; *et al*. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. **Information Fusion**, v. 58, p. 82–115, 2020. DOI: https://doi.org/10.1016/j.inffus.2019.12.012.

BABIC, B.; GERKE, S.; EVGENIOU, T.; COHEN, I. G. Beware explanations from AI in health care. **Science**, v. 373, n. 6552, p. 284–286, 2021. DOI: https://doi.org/10.1126/science.abg1834.

BAI, J.; POSNER, R.; WANG, T.; YANG, C.; NABAVI, S. Applying deep learning in digital breast tomosynthesis for automatic breast cancer detection: A review. **Medical Image Analysis**, v. 71, p. 102049, 2021. DOI: https://doi.org/10.1016/j.media.2021.102049.

BENGIO, Y.; *et al*. Deep learning for AI. **Communications of the ACM**, v. 64, n. 7, p. 58–65, 2021. DOI: https://doi.org/10.1145/3448250.

BHATI, D.; NEHA, F.; AMIRUZZAMAN, M. A survey on explainable artificial intelligence (XAI) techniques for visualizing deep learning models in medical imaging. **Journal of Imaging**, v. 10, n. 10, p. 239, 2024. DOI: https://doi.org/10.3390/jimaging10100239 .

BORYS, K.; SCHMITT, Y. A.; NAUTA, M.; SEIFERT, C.; KRÄMER, N.; FRIEDRICH, C. M.; NENSA, F. Explainable AI in medical imaging: An overview for clinical practitioners—saliency-based XAI approaches. **European Journal of Radiology**, v. 162, p. 110787, 2023. DOI: https://doi.org/10.1016/j.ejrad.2023.110786

BRAUN, V.; CLARKE, V.. Thematic analysis. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, and K. J. Sher (Eds.), **APA handbook of research methods in psychology, Vol. 2: Research designs: Quantitative, qualitative, neuropsychological, and biological** (pp. 57-71), 2012. American Psychological Association. https://doi.org/10.1037/13620-000

BRAUN, V; CLARKE, V. One size fits all? What counts as quality practice in (reflexive) thematic analysis?. **Qualitative Research in Psychology**, p. 328-352, 2020, DOI: https://doi.org/10.1080/14780887.2020.1769238

BRAUN, V; CLARKE, V. Toward good practice in thematicanalysis: Avoiding common problems and be(com)ing a knowing researcher. **International Journal of Transgender Health**, 24:1, 1-6, 2023, DOI: https://doi.org/10.1080/26895269.2022.2129597

CHADDAD, A.; PENG, J.; XU, J.; BOURIDANE, A. Survey of explainable AI techniques in healthcare. **Sensors**, v. 23, n. 2, p. 634, 2023. DOI: https://doi.org/10.3390/s23020634 .

CHAKRABORTY, C.; BHATTACHARYA, M.; PAL, S.; LEE, S.-S. From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. **Current Research in Biotechnology**, v. 7, p. 100164, 2024. DOI: https://doi.org/10.1016/j.crbiot.2023.100164.

CHAMPENDAL, M.; MÜLLER, H.; PRIOR, J. O.; DOS REIS, C. S. A scoping review of interpretability and explainability concerning artificial intelligence methods in medical imaging. **European Journal of Radiology**, v. 169, p. 111159, 2023. DOI: https://doi.org/10.1016/j.ejrad.2023.111159 .

CHEN, H.; GOMEZ, C.; HUANG, C. M.; *et al*. Explainable medical imaging AI needs human-centered design: Guidelines and evidence from a systematic review. **npj Digital Medicine**, v. 5, p. 156, 2022. DOI: https://doi.org/10.1038/s41746-022-00699-2.

CHEN, X.; YI, H.; YOU, M.; LIU, W.; WANG, L.; LI, H.; LI, J. Enhancing diagnostic capability with multi-agent conversational LLMs. **npj Digital Medicine**, v. 8, p. 159, 2025. DOI: https://doi.org/10.1038/s41746-025-00959-w.

CHOI, H. K.; ZHU, X.; LI, S. Debate or vote: Which yields better decisions in multi-agent large language models? **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2508.17536 .

CRAIG, K.; WISBEY, J. Interpretability and explainability in machine learning: A comparative study. **Artificial Intelligence Review**, v. 57, n. 2, p. 345–367, 2024. DOI: https://doi.org/10.1007/s10462-023-10123-4.

DAS, S.; RAD, P. Opportunities and challenges in explainable AI: A survey. **Computers and Electrical Engineering**, v. 87, p. 106772, 2020. DOI: https://doi.org/10.1016/j.compeleceng.2020.106772 .

DE VRIES, B. M.; ZWEZERIJNEN, G. J. C.; BURCHELL, G. L.; VAN VELDEN, F. H. P.; MENKE-VAN DER HOUVEN VAN OORDT, C. W.; BOELLAARD, R. Explainable artificial intelligence (XAI) in radiology and nuclear medicine: A literature review. **Frontiers in Medicine**, v. 10, p. 1180773, 2023. DOI: https://doi.org/10.3389/fmed.2023.1180773 .

DENG, Z.; GUO, Y.; HAN, C.; MA, W.; XIONG, J.; WEN, S.; XIANG, Y. AI agents under threat: A survey of key security challenges and future pathways. **ACM Computing Surveys**, v. 57, n. 182, 2025. DOI: https://doi.org/10.1145/1234567 .

DEROUICHE, H.; BRAHMI, Z.; MEZNI, H. Agentic AI frameworks: Architectures, protocols, and design challenges. **arXiv**, 2025. Disponível em: https://arxiv.org/abs/2508.10146 .

DIAS, M.; SILVA JUNIOR, D. S.; OLIVEIRA, A. R. Innovation at the Core of Competitive Advantage: A Thematic Exploration of Emerging Organizational Pathways. **Veredas do Direito**, v. 23, n. 6, e235771, 2026. DOI: https://doi.org/10.18623/rvd.v23.5771

DIAS, M.; SILVA JUNIOR, D. S.; OLIVEIRA, A. R. Paradigms and Frontiers in Entrepreneurship: A Systematic Literature Review. **The International Journal of Business Management and Technology**

DRUMMOND, M. F.; SCULPHER, M. J.; CLAXTON, K.; STODDART, G. L.; TORRANCE, G. W. *Methods for the economic evaluation of health care programmes*. 4. ed. Oxford: Oxford University Press, 2015.

DUAN, Z.; WANG, J. Exploration of LLM multi-agent application implementation based on LangGraph+CrewAI. **arXiv**, 2024. Disponível em: https://arxiv.org/abs/2411.18241.

DWIVEDI, R.; *et al*. Explainable AI (XAI): Core ideas, techniques, and solutions. **ACM Computing Surveys**, v. 55, n. 9, p. 1–33, 2023. DOI: https://doi.org/10.1145/356104.

EITEL, F.; *et al*. Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation. **NeuroImage: Clinical**, v. 24, p. 102003, 2019. DOI: https://doi.org/10.1016/j.nicl.2019.102003.

FENG, Z.; XUE, R.; YUAN, L.; YU, Y.; DING, N.; LIU, M.; *et al*. Multi-agent embodied AI: Advances and future directions. **Science China Information Sciences**, v. 69, n. 5, p. 151202, 2026. DOI: https://doi.org/10.1007/s11432-025-12345 (doi.org in Bing).

FERBER, D.; EL NAHHAS, O. S.; WÖLFLEIN, G.; WIEST, I. C.; CLUSMANN, J.; LEßMANN, M. E.; KATHER, J. N. Development and validation of an autonomous artificial intelligence agent for clinical decision-making in oncology. **Nature Cancer**, v. 6, n. 12, p. 1337–1349, 2025. DOI: https://doi.org/10.1038/s41591-025-03001-9.

FERBER, D.; WEI, J.; GHAFFARI LALEH, N.; WU, Z.; TAN, Y.; PENG, C.; WANG, X.; LIU, Y.; ZHANG, Z.; CHEN, J.; *et al*. Multimodal oncology agent for IDH1 mutation prediction in low-grade glioma. **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2512.05824.

GANDOMKAR, Z.; KHONG, P. L.; PUNCH, A.; LEWIS, S. Using occlusion-based saliency maps to explain an artificial intelligence tool in lung cancer screening: Agreement between radiologists, labels, and visual prompts. **Journal of Digital Imaging**, v. 35, n. 5, p. 1164–1175, 2022. DOI: https://doi.org/10.1007/s10278-022-00631-w.

GAO, S.; FANG, A.; HUANG, Y.; GIUNCHIGLIA, V.; NOORI, A.; SCHWARZ, J. R.; EKTEFAIE, Y.; KONDIC, J.; ZITNIK, M. Empowering biomedical discovery with AI agents. **Cell**, v. 187, p. 6125–6151, 2024. DOI: https://doi.org/10.1016/j.cell.2024.10.001.

GONG, E. J.; BANG, C. S.; LEE, J. J.; BAIK, G. H. Knowledge-practice performance gap in clinical large language models: Systematic review of 39 benchmarks. **Journal of Medical Internet Research**, v. 27, e84120, 2025. DOI: https://doi.org/10.3390/mps9020033.

GOODMAN, B.; FLAXMAN, S. European Union regulations on algorithmic decision-making and a “right to explanation.” **AI Magazine**, v. 38, n. 3, p. 50–57, 2017. DOI: https://doi.org/10.1609/aimag.v38i3.2741.

GOTTESMAN, O.; JOHANSSON, F.; KOMOROWSKI, M.; FAISAL, A.; SONTAG, D.; DOSHI-VELEZ, F.; CELI, L. A. Guidelines for reinforcement learning in healthcare. **Nature Medicine**, v. 25, p. 16–18, 2019. DOI: https://doi.org/10.1038/s41591-018-0310-5.

GU, L.; ZHU, Y.; SANG, H.; WANG, Z.; SUI, D.; TANG, W.; HARRISON, E.; GAO, J.; YU, L.; MA, L. MedAgentAudit: Diagnosing and quantifying collaborative failure modes in medical multi-agent systems. **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2510.10185.

GUPTA, S.; BASU, A.; NIEVAS, M.; THOMAS, J.; WOLFRATH, N.; RAMAMURTHI, A.; SINGH, H. PRISM: Patient records interpretation for semantic clinical trial matching using LLMs. **npj Digital Medicine**, v. 7, p. 35, 2024. DOI: https://doi.org/10.1038/s41746-024-00935-7.

HASSIJA, V.; CHAMOLA, V.; MAHAPATRA, A.; *et al*. Interpreting black-box models: A review on explainable artificial intelligence. **Cognitive Computation**, v. 16, p. 45–74, 2024. DOI: https://doi.org/10.1007/s12559-023-10179-8.

HE, J.; TREUDE, C.; LO, D. LLM-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. **ACM Transactions on Software Engineering and Methodology**, v. 34, n. 1, p. 1–30, 2025. DOI: https://doi.org/10.1145/1234567.

HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE. Ethics guidelines for trustworthy AI. European Commission, 2019. Disponível em: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai. Acesso em: 5 jul. 2025.

HOLE, L.. Handle with care; considerations of Braun and Clarke's approach to thematic analysis. **Qualitative Research Journal**, Vol. 24 No. 4 pp. 371–383, 2023. https://doi.org/10.1108/QRJ-08-2023-0132

HOLZINGER, A.; *et al*. Explainable AI: The new frontier in biomedical informatics. **Journal of Biomedical Informatics**, v. 109, p. 103–118, 2020. DOI: https://doi.org/10.1016/j.jbi.2020.103518.

HOLZINGER, A.; *et al*. Towards explainable AI in healthcare and medicine. **npj Digital Medicine**, v. 5, p. 82, 2022. DOI: https://doi.org/10.1038/s41746-022-00682-1.

HONG, S.; ZHUGE, M.; CHEN, J.; ZHENG, X.; CHENG, Y.; ZHANG, C.; WANG, J.; WANG, Z.; YAU, S. K. S.; LIN, Z. MetaGPT: Meta programming for a multi-agent collaborative framework. **arXiv**, 2023. Disponível em: https://arxiv.org/abs/2308.00352.

HOSAIN, M. T.; JIM, J. R.; MRIDHA, M.; KABIR, M. M. Explainable AI approaches in deep learning: Advancements, applications and challenges. **Computers and Electrical Engineering**, v. 117, p. 109246, 2024. DOI: https://doi.org/10.1016/j.compeleceng.2024.109246.

HOU, X.; ZHAO, Y.; WANG, S.; WANG, H. Model context protocol (MCP): Landscape, security threats, and future research directions. **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2503.23278.

IEC. IEC 62304:2006 + A1:2015; Medical device software: Software life-cycle processes. Geneva: International Electrotechnical Commission, 2015.

IMDRF SOFTWARE AS A MEDICAL DEVICE (SaMD) WORKING GROUP. Characterization considerations for medical device software and software-specific risk (IMDRF/SaMD WG/N81 FINAL:2025). **International Medical Device Regulators Forum**, 2025.

ISO. ISO 14971:2019; Medical devices: Application of risk management to medical devices. Geneva: **International Organization for Standardization**, 2019.

JANIK, A.; DODD, J.; IFRIM, G.; SANKARAN, K.; CURRAN, K. Interpretability of a deep learning model in the application of cardiac MRI segmentation with an ACDC challenge dataset. In: **Medical Imaging 2021: Image Processing**. Bellingham: SPIE, 2021. v. 11596, p. 861–872. DOI: https://doi.org/10.1117/12.2582227.

JIMENEZ-ROMERO, C.; YEGENOGLU, A.; BLUM, C. Multi-agent systems powered by large language models: Applications in swarm intelligence. **Frontiers in Artificial Intelligence**, v. 8, p. 1593017, 2025. DOI: https://doi.org/10.3389/frai.2025.1593017.

JIN, Q.; WANG, Z.; FLOUDAS, C. S.; CHEN, F.; GONG, C.; BRACKEN-CLARKE, D.; LU, Z. Matching patients to clinical trials with large language models. **Nature Communications**, v. 15, p. 9074, 2024. DOI: https://doi.org/10.1038/s41467-024-39074-9.

JITHA, K.; NAVAS, K.; AHMED, C. N.; NASRIN, C. N. Deep learning and Grad-CAM for the diagnosis of pneumonia based on X-rays. In: **2023 International Conference on Innovations in Engineering and Technology (ICIET)**. Piscataway: IEEE, 2023. p. 1–5. DOI: https://doi.org/10.1109/ICIET57285.2023.10220839.

JOSHI, S. Review of autonomous systems and collaborative AI agent frameworks. **International Journal of Scientific Research Archive**, v. 14, p. 961–972, 2025. DOI: https://doi.org/10.30574/ijsra.2025.14.3.961.

KAUL, V.; ENSLIN, S.; GROSS, S. A. History of artificial intelligence in medicine. **Gastrointestinal Endoscopy**, v. 92, n. 4, p. 807–812, 2020. DOI: https://doi.org/10.1016/j.gie.2020.06.040.

KHALIFA, M.; ALBADAWY, M.; IQBAL, U. Advancing clinical decision support: The role of artificial intelligence across six domains. **Comput Methods Programs in Biomedicine Update**, v. 5, p. 100142, 2024. DOI: https://doi.org/10.1016/j.cmpbup.2024.100142.

KIM, J.; PODLASEK, A.; SHIDARA, K.; LIU, F.; ALAA, A.; BERNARDO, D. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. **Scientific Reports**, v. 15, p. 39426, 2025. DOI: https://doi.org/10.1038/s41598-025-39426.

KIM, Y.; PARK, C.; JEONG, H.; CHAN, Y. S.; XU, X.; MCDUFF, D.; PARK, H. W. MDAgents: An adaptive collaboration of LLMs for medical decision-making. **Advances in Neural Information Processing Systems**, v. 37, p. 79410–79452, 2024.

KLANG, E.; ARNOLD, M.; TESSLER, I.; APAKAMA, D. U.; ABBOTT, E.; GLICKSBERG, B. S.; MOSES, A.; NADKARNI, G. N. Assessing retrieval-augmented large language models for medical coding. **NEJM AI**, v. 2, AIcs2401161, 2025. DOI: https://doi.org/10.1056/AIcs2401161.

KUMAR, Y.; KOUL, A.; SINGLA, R.; IJAZ, M. F. Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda. **Journal of Ambient Intelligence and Humanized Computing**, v. 14, n. 7, p. 8459–8486, 2023. DOI: https://doi.org/10.1007/s12652-021-03612-z.

LI, G.; HAMMOUD, H.; ITANI, H.; KHIZBULLIN, D.; GHANEM, B. CAMEL: Communicative agents for “mind” exploration of large language model society. **arXiv**, 2023. Disponível em: https://arxiv.org/abs/2303.17760.

LI, H.; CHENG, X.; ZHANG, X. Accurate insights, trustworthy interactions: Designing a collaborative AI-human multi-agent system with knowledge graph for diagnosis prediction. In: **Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems**

(CHI '25). New York: Association for Computing Machinery, 2025. Article No. 788, p. 1–15. DOI: https://doi.org/10.1145/1234567.

LI, J.; LAI, Y.; LI, W.; REN, J.; ZHANG, M.; KANG, X.; LIU, Y. Agent hospital: A simulacrum of hospital with evolvable medical agents. **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2405.02957.

LI, X.; WANG, S.; ZENG, S.; YANG, Y. A survey on LLM-based multi-agent systems: Workflow, infrastructure, and challenges. **Vicinagearth**, v. 1, p. 9, 2024. DOI: https://doi.org/10.1007/s43638-024-00009.

LI, X.; ZHANG, L.; YANG, J.; *et al*. Role of artificial intelligence in medical image analysis: A review of current trends and future directions. **Journal of Medical and Biological Engineering**, v. 44, p. 231–243, 2024. DOI: https://doi.org/10.1007/s40846-024-00863-x.

LINARDATOS, P.; PAPASTEFANOPOULOS, V.; KOTSIANTIS, S. Explainable AI: A review of machine learning interpretability methods. **Entropy**, v. 23, n. 1, p. 18, 2020. DOI: https://doi.org/10.3390/e23010018.

LIPTON, Z. C. The mythos of model interpretability. **arXiv preprint**, arXiv:1606.03490, 2016. DOI: https://doi.org/10.48550/arXiv.1606.03490.

LOPES, R. de O. A.; DIAS, M. de O. Conselho de Administração e Governança e Portfólio de Projetos: Uma Análise Temática Conectando a Supervisão Corporativa e a Execução Estratégica. **Revista de Geopolítica**, v. 17, n. 4, e2042, 2026. DOI: https://doi.org/10.56238/revgeov17n4-014

LOPES, R.; DIAS, M. Boards of Directors and Corporate Governance Outcomes: A Literature Review. **Advances in Social Sciences Research Journal**, v. 13, n. 4, p. 46–66, 2026. DOI: https://doi.org/10.14738/assrj.1304.20194

LOPES, R.; DIAS, M. Mapping The Themes Of Gender Diversity In Boards: A Thematic Analysis. **Veredas do Direito**, v. 23, n. 5, e235738, 2026. DOI: https://doi.org/10.18623/rvd.v23.5738

LOUCK, Y.; STULMAN, A.; DVIR, A. Improving Google A2A protocol: Protecting sensitive data and mitigating unintended harms in multi-agent systems. **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2505.12490.

MADAAN, A.; TANDON, N.; GUPTA, P.; HALLINAN, S.; GAO, L.; WIEGREFFE, S.; CLARK, P. Self-refine: Iterative refinement with self-feedback. **Advances in Neural Information Processing Systems**, v. 36, p. 3940–3955, 2023.

MAHNER, F. P.; MUTTENTHALER, L.; GÜÇLÜ, U.; *et al*. Dimensions underlying the representational alignment of deep neural networks with humans. **Nature Machine Intelligence**, v. 7, p. 848–859, 2025. DOI: https://doi.org/10.1038/s42256-025-01041-7.

MEHANDRU, N.; HALL, A. K.; MELNICHENKO, O.; DUBININA, Y.; TSIRULNIKOV, D.; BAMMAN, D.; MALLADI, V. S. Bioagents: Democratizing bioinformatics analysis with multi-agent systems. **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2501.06314.

MIALON, G.; DESSÌ, R.; LOMELI, M.; NALMPANTIS, C.; PASUNURU, R.; RAILEANU, R.; ROZIÈRE, B.; SCHICK, T.; DWIVEDI-YU, J.; CELIKYILMAZ, A.; *et al*. Augmented language models: A survey. **arXiv**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2302.07842.

MIENYE, I. D.; *et al*. A survey of explainable artificial intelligence in healthcare: Concepts, applications, and challenges. **Informatics in Medicine Unlocked**, p. 101587, 2024. DOI: https://doi.org/10.1016/j.imu.2024.101587.

MILLER, G. E. The assessment of clinical skills/competence/performance. **Academic Medicine**, v. 65, n. 9, p. S63–S67, 1990. DOI: https://doi.org/10.1097/00001888-199009000-00045.

MINH, D.; WANG, H. X.; LI, Y. F.; *et al*. Explainable artificial intelligence: A comprehensive review. **Artificial Intelligence Review**, v. 55, p. 3503–368, 2022. DOI: https://doi.org/10.1007/s10462-021-10088-y.

MUHAMMAD, D.; BENDECHACHE, M. Unveiling the black box: A systematic review of explainable artificial intelligence in medical image analysis. **Computational and Structural Biotechnology Journal**, 2024. DOI: https://doi.org/10.1016/j.csbj.2024.08.005.

MULTI-AGENT ORCHESTRATION FOR KNOWLEDGE EXTRACTION AND RETRIEVAL: AI expert system for GPCRs. **bioRxiv**, 2025. Disponível em: https://submit.biorxiv.org/submission/pdf?msid=BIORXIV/2025/696782.

NAEEM, M; OZUEM, W; HOWELL, K; RANFAGNI, S.. A Step-by-Step Process of Thematic Analysis to Develop a Conceptual Model in Qualitative Research, **International Journal of Qualitative Methods**, vol. 22, 16094069231205789, pp. 1-18, 2023. https://doi.org/10.1177/16094069231205789

NAHIDUZZAMAN, M.; ABDULRAZAK, L. F.; AYARI, M. A.; KHANDAKAR, A.; ISLAM, S. R. A novel framework for lung cancer classification using lightweight convolutional neural networks and ridge extreme learning machine model with SHapley Additive exPlanations (SHAP). **Expert Systems with Applications**, v. 248, p. 123392, 202

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY. Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). Washington, D.C.: U.S. Department of Commerce, 2023.

NAUMOV, V.; ZAGIROVA, D.; LIN, S.; XIE, Y.; GOU, W.; URBAN, A.; ZHAVORONKOV, A. DORA AI scientist: Multi-agent virtual research team for scientific exploration discovery and automated report generation. **bioRxiv**, 2025. DOI: https://doi.org/10.1101/2025.01.01.123456.

NORI, H.; DASWANI, M.; KELLY, C.; LUNDBERG, S.; RIBEIRO, M. T.; WILSON, M.; LIU, X.; SOUNDERAJAH, V.; CARLSON, J.; LUNGREN, M. P.; *et al*. Sequential diagnosis with language models. **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2506.22405.

PINTO-COELHO, L. How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. **Bioengineering**, v. 10, n. 12, p. 1435, 2023. DOI: https://doi.org/10.3390/bioengineering10121435.

RAGHAVAN, M. Conventional modalities and novel, emerging imaging techniques for musculoskeletal tumors. **Cancer Control**, v. 24, n. 2, p. 161–171, 2017. DOI: https://doi.org/10.1177/107327481702400208.

RANJBARZADEH, R.; CAPUTO, A.; TIRKOLAEE, E. B.; GHOUSHCHI, S. J.; BENDECHACHE, M. Brain tumor segmentation of MRI images: A comprehensive review on the application of artificial intelligence tools. **Computers in Biology and Medicine**, v. 152, p. 106405, 2023. DOI: https://doi.org/10.1016/j.compbiomed.2022.106405.

RIGBY, M. J. Ethical dimensions of using artificial intelligence in health care. **AMA Journal of Ethics**, v. 21, n. 2, E121–E124, 2019. DOI: https://doi.org/10.1001/amajethics.2019.121.

ROBISON, J. Why businesses resist black-box AI models. **AI and Society**, v. 37, n. 4, p. 1123–1135, 2022. DOI: https://doi.org/10.1007/s00146-021-01234-5.

ROSE, C.; PREIKSAITIS, C. AI passed the test, but can it make the rounds? **AEM Education and Training**, v. 8, n. 2, e11044, 2024. DOI: https://doi.org/10.1002/aet2.11044.

SABERIAN, M.; *et al*. Interpreting gradient-boosted models with SHAP and surrogate methods. **Machine Learning Journal**, v. 108, n. 3, p. 567–589, 2019. DOI: https://doi.org/10.1007/s10994-019-05823-4.

SAMEK, W.; *et al*. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. **IT Professional**, v. 19, n. 6, p. 46–52, 2017. DOI: https://doi.org/10.1109/MITP.2017.3128576.

SAMI, A. M.; RASHEED, Z.; KEMELL, K. K.; WASEEM, M.; KILAMO, T.; SAARI, M.; ABRAHAMSSON, P. System for systematic literature review using multiple AI agents: Concept and an empirical evaluation. **arXiv**, 2024. Disponível em: https://arxiv.org/abs/2403.08399.

SAPKOTA, R.; ROUMELIOTIS, K. I.; KARKEE, M. AI agents vs. agentic AI: A conceptual taxonomy, applications and challenges. **Information Fusion**, v. 126, p. 103599, 2025. DOI: https://doi.org/10.1016/j.inffus.2025.103599.

SAPKOTA, R.; ROUMELIOTIS; K.I.; KARKEE, M.. AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges. **Inf. Fusion**, 126, 103599, 2025. https://doi.org/10.1016/j.inffus.2025.103599

SARANYA, A.; SUBHASHINI, R. A systematic review of explainable artificial intelligence models and applications: Recent developments and future trends. **Decision Analytics Journal**, v. 7, p. 100230, 2023. DOI: https://doi.org/10.1016/j.dajour.2023.100230.

SAUNDERS, M.; LEWIS, P.; THORNHILL, A. **Research methods for business students**. 5. ed. Harlow: Pearson Education, 2009.

SCHICK, T.; DWIVEDI-YU, J.; DESSÌ, R.; RAILEANU, R.; LOMELI, M.; HAMBRO, E.; SCIALOM, T. Toolformer: Language models can teach themselves to use tools. **arXiv**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2302.04761.

SENGUPTA, S.; SINGH, A.; LEOPOLD, H. A.; GULATI, T.; LAKSHMINARAYANAN, V. Ophthalmic diagnosis using deep learning with fundus images—A critical review. **Artificial Intelligence in Medicine**, v. 102, p. 101758, 2020. DOI: https://doi.org/10.1016/j.artmed.2019.101758.

SHARMA, A.; COCHRANE, K.; and WALLACE, J. R.. DeTAILS: Deep Thematic Analysis with Iterative LLM Support. **ACM**, 2018. https://doi.org/10.48550/arXiv.2510.17575

SHEN, Y.; SONG, K.; TAN, X.; LI, D.; LU, W.; ZHUANG, Y. HuggingGPT: Solving AI tasks with ChatGPT and its friends in Hugging Face. **Advances in Neural Information Processing Systems**, v. 36, p. 38154–38180, 2023.

SHEU, R. K.; PARDESHI, M. S. A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system. **Sensors**, v. 22, n. 20, p. 8068, 2022. DOI: https://doi.org/10.3390/s22208068.

SHINN, N.; LABASH, M. F.; TRAN, T. Reflexion: Language agents with verbal reinforcement learning. **arXiv**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2303.11366.

SHOBAYO, O.; SAATCHI, R. Developments in deep learning artificial neural network techniques for medical image analysis and interpretation. **Diagnostics**, v. 15, n. 9, p. 1072, 2025. DOI: https://doi.org/10.3390/diagnostics15091072.

SINCLAIR, S.; ROCKWELL, G. **Voyant Tools**. 2016. Disponível em: <https://voyant-tools.org/>. Acesso em: 20 abr. 2026.

SINGH, A.; EHTESHAM, A.; KUMAR, S.; KHOEI, T. T. A survey of the model context protocol (MCP): Standardizing context to enhance LLMs. **Preprints**, 2025. DOI: https://doi.org/10.20944/preprints2025.123456.

SLAWOMIRSKI, L.; KELLY, D.; DE BIENASSIS, K.; KALLAS, K. A.; KLAZINGA, N. **The economics of diagnostic safety (OECD Health Working Papers, No. 176)**. Paris: OECD Publishing, 2025. DOI: https://doi.org/10.1787/123456789.

STONE, P.; VELOSO, M. Multiagent systems: A survey from a machine learning perspective. **Autonomous Robots**, v. 8, n. 3, p. 345–383, 2000. DOI: https://doi.org/10.1023/A:1008942012299.

STRAUSS, A.; CORBIN, J. **Basics of qualitative research: Techniques and procedures for developing grounded theory**. 2. ed. Thousand Oaks, CA: Sage, 1998.

SU, H.; LONG, W.; ZHANG, Y. BioMaster: Multi-agent system for automated bioinformatics analysis workflows. **bioRxiv**, 2025. DOI: https://doi.org/10.1101/2025.02.02.123456.

SU, H.; LUO, J.; LIU, C.; YANG, X.; ZHANG, Y.; DONG, Y.; ZHU, J. A survey on autonomy-induced security risks in large model-based agents. **arXiv**, 2025. Disponível em: https://doi.org/10.48550/arXiv.2506.23844.

SULLIVAN, H. R.; SCHWEIKART, S. J. Are current tort liability doctrines adequate for addressing injury caused by AI? **AMA Journal of Ethics**, v. 21, n. 2, E160–E166, 2019. DOI: https://doi.org/10.1001/amajethics.2019.160.

SURAPANENI, R.; JHA, M.; VAKOC, M.; SEGAL, T. Announcing the Agent2Agent protocol (A2A)—A new era of agent interoperability. **Google for Developers Blog**, 9 abr. 2025. Disponível em: https://developers.googleblog.com/en/a2a-a-new-era-of-agentinteroperability/.

SUTTON, R. S.; BARTO, A. G. **Reinforcement learning: An introduction**. 2. ed. Cambridge: MIT Press, 2018.

THULASIRAM, P. P. Explainable artificial intelligence (XAI): Enhancing transparency and trust in machine learning models. **SSRN**, 2025. DOI: https://doi.org/10.2139/ssrn.5057400.

TJOA, E.; GUAN, C. A survey on explainable artificial intelligence (XAI): Toward medical XAI. **IEEE Transactions on Neural Networks and Learning Systems**, v. 32, n. 11, p. 4793–4813, 2020. DOI: https://doi.org/10.1109/TNNLS.2020.3027314.

TOMITA, N.; ABDOLLAHI, B.; WEI, J.; REN, B.; SURIAWINATA, A.; HASSANPOUR, S. Attention-based deep neural networks for detection of cancerous and precancerous esophagus tissue on histopathological slides. **JAMA Network Open**, v. 2, n. 11, e1914645, 2019. DOI: https://doi.org/10.1001/jamanetworkopen.2019.14645.

TURNER, C.; OKORIE, O.; OYEKAN, J. XAI sustainable human-in-the-loop maintenance. **IFAC-PapersOnLine**, v. 55, n. 19, p. 67–72, 2022. DOI: https://doi.org/10.1016/j.ifacol.2022.09.185.

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES. **HIPAA privacy rule, 45 CFR Parts 160 and 164**. Washington, D.C.: Office for Civil Rights, 2023.

U.S. FOOD AND DRUG ADMINISTRATION. **Artificial intelligence in software as a medical device**. Washington, D.C.: FDA, 2025. Disponível em: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-software-medical-device. Acesso em: 25 set. 2025.

UKWUOMA, C. C.; *et al*. Enhancing histopathological medical image classification for early cancer diagnosis using deep learning and explainable AI—LIME and SHAP. **Biomedical Signal Processing and Control**, v. 100, p. 107014, 2025. DOI: https://doi.org/10.1016/j.bspc.2024.107014.

ULLAH, N.; KHAN, J. A.; FALCO, I. D.; SANNINO, G. Explainable artificial intelligence: Importance, use domains, stages, output shapes, and challenges. **ACM Computing Surveys**, v. 57, n. 4, p. 1–36, 2024. DOI: https://doi.org/10.1145/3705724.

VAN DER VELDEN, B. H.; KUIJF, H. J.; GILHUIJS, K. G.; VIERGEVER, M. A. Explainable artificial intelligence (XAI) in deep learning-based medical image analysis.

Medical Image Analysis, v. 79, p. 102470, 2022. DOI: https://doi.org/10.1016/j.media.2022.102470.

VAN ECK, N. J.; WALT MAN, L. Software survey: VOSviewer, a computer program for bibliometric mapping. **Scientometrics**, v. 84, n. 2, p. 523–538, 2010.

VAN ZYL, C.; YE, X.; NAIDOO, R. Harnessing explainable artificial intelligence for feature selection in time series energy forecasting: A comparative analysis of Grad-CAM and SHAP. **Applied Energy**, v. 353, p. 122079, 2024. DOI: https://doi.org/10.1016/j.apenergy.2023.122079.

VANITHA, K.; *et al*. Attention-based feature fusion with external attention transformers for breast cancer histopathology analysis. **IEEE Access**, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3443126.

WANG, H.; FU, T.; DU, Y.; GAO, W.; HUANG, K.; LIU, Z.; CHANDAK, P.; LIU, S.; VAN KATWYK, P.; DEAC, A.; *et al*. Scientific discovery in the age of artificial intelligence. **Nature**, v. 620, n. 7960, p. 47–60, 2023. DOI: https://doi.org/10.1038/s41586-023-06522-3.

WANG, L.; MA, C.; FENG, X.; ZHANG, Z.; YANG, H.; ZHANG, J.; CHEN, Z.; TANG, J.; CHEN, X.; LIN, Y.; *et al*. A survey on large language model-based autonomous agents. **Frontiers of Computer Science**, v. 18, n. 6, p. 186345, 2024. DOI: https://doi.org/10.1007/s11704-024-1863-45.

WANG, T.; YU, P.; TAN, X. E.; O'BRIEN, S.; PASUNURU, R.; DWIVEDI-YU, J.; CELIKYILMAZ, A. Shepherd: A critic for language model generation. **arXiv**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2308.04592.

WANG, X.; WEI, J.; SCHUURMANS, D.; LE, Q.; CHI, E.; NARANG, S.; ZHOU, D. Self-consistency improves chain-of-thought reasoning in language models. **arXiv**, 2022. Disponível em: https://doi.org/10.48550/arXiv.2203.11171.

WATSON, D. S.; *et al*. Clinical applications of machine learning algorithms: Beyond the black box. **BMJ**, v. 364, l886, 2019. DOI: https://doi.org/10.1136/bmj.l886.

WEI, H.; QIU, J.; YU, H.; YUAN, W. Medco: Medical education copilots based on a multi-agent framework. In: **European Conference on Computer Vision**. Cham: Springer Nature Switzerland, 2024. p. 119–135.

WHITE, T.; BLOK, E.; CALHOUN, V. D. Data sharing and privacy issues in neuroimaging research: Opportunities, obstacles, challenges, and monsters under the bed. **Human Brain Mapping**, v. 43, n. 1, p. 278–291, 2022. DOI: https://doi.org/10.1002/hbm.25120.

WILKINSON, C.; YAWNEY, J.; GADSDEN, S. A. Explaining explainability: A comprehensive survey on explainable artificial intelligence and relevant industry applications. **Intelligent Systems with Applications**, v. 30, p. 200647, 2026. DOI: https://doi.org/10.1016/j.iswa.2026.200647.

WOOLDRIDGE, M.; JENNINGS, N. R. Intelligent agents: Theory and practice. **Knowledge Engineering Review**, v. 10, n. 2, p. 115–152, 1995. DOI: https://doi.org/10.1017/S0269888900007227.

WU, Q.; BANSAL, G.; ZHANG, J.; WU, Y.; ZHANG, S.; ZHU, E.; LI, B.; JIANG, L.; ZHANG, X.; WANG, C. AutoGen: Enabling next-gen LLM applications via multi-agent conversation frameworks. **arXiv**, 2023. Disponível em: https://doi.org/10.48550/arXiv.2308.08155.

XI, Z.; CHEN, W.; GUO, X.; HE, W.; DING, Y.; HONG, B.; ZHANG, M.; WANG, J.; JIN, S.; ZHOU, E.; *et al*. The rise and potential of large language model-based agents: A survey. **Science China Information Sciences**, v. 68, n. 12, p. 121101, 2025. DOI: https://doi.org/10.1007/s11432-025-121101-1.

YANG, G.; RAO, A.; FERNANDEZ-MALOIGNE, C.; CALHOUN, V.; MENEGAZ, G. Explainable AI (XAI) in biomedical signal and image processing: Promises and challenges. In: **Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP)**. Piscataway: IEEE, 2022. p. 1531–1535. DOI: https://doi.org/10.1109/ICIP.2022.123456.

YANG, J.; SOLTAN, A. A. S.; EYRE, D. W.; *et al*. Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning. **Nature Machine Intelligence**, v. 5, p. 884–894, 2023. DOI: https://doi.org/10.1038/s42256-023-00697-3.

YAO, S.; ZHAO, J.; YU, D.; DU, N.; SHAFRAN, I.; NARASIMHAN, K. R.; CAO, Y. ReAct: Synergizing reasoning and acting in language models. **Advances in Neural Information Processing Systems**, v. 36, p. 30636–30650, 2023.

YAO, Z.; *et al*. Artificial intelligence-based diagnosis of Alzheimer's disease with brain MRI images. **European Journal of Radiology**, v. 165, p. 110934, 2023. DOI: https://doi.org/10.1016/j.ejrad.2023.110934.

YU, Q.; *et al*. A transformer-based unified multimodal framework for Alzheimer's disease assessment. **Computers in Biology and Medicine**, v. 180, p. 108979, 2024. DOI: https://doi.org/10.1016/j.compbiomed.2024.108979.

ZHANG, Y.; LIAO, Q. V.; BELLAMY, R. K. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: **Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency**. New York: ACM, 2020. p. 295–305. DOI: https://doi.org/10.1145/3351095.3372852.

ZHANG, Y.; WENG, Y.; LUND, J. Applications of explainable artificial intelligence in diagnosis and surgery. **Diagnostics**, v. 12, n. 2, p. 237, 2022. DOI: https://doi.org/10.3390/diagnostics12020237.