v. 05, nº 02 - jul/dec 2025
ISSN 2763-8685

LATIN AMERICAN JOURNAL OF EUROPEAN STUDIES

BRIDGE WATCH
JEAN MONNET POLICY DEBATE

LACES
Latin American Center of European Studies

Co-funded by the European Union

TABLE OF CONTENTS

https://dx.doi.org/10.51799/2763-8685v5n2008

ARTIFICIAL INTELLIGENCE:

challenges of explainability on disinformation through chatbots[1]

René Palacios Garita[2]

ABSTRACT: The paper analyses the challenges of explainability of artificial intelligence, particularly in chatbots, focusing on the risks of misinformation, polarization and algorithmic discrimination. The evolution of chatbots is explored, from rule-based systems to advanced language models such as A.L.I.C.E.®, Replika®, ChatGPT®, Bard® and DeepSeek®. Furthermore, Explained Artificial Intelligence (XAI) is proposed as a solution to mitigate these risks through explainability of AI systems. Finally, the application of XAI principles in different chatbots is evaluated, identifying their strengths and weaknesses in terms of explainability, interpretability, and ethics. The paper concludes by highlighting the importance of XAI for responsible and ethical use of AI.

KEYWORDS: Explainable Artificial Intelligence. Disinformation. Risks. Chatbots. Machine learning.

INTELIGENCIA ARTIFICIAL: RETOS DE LA EXPLICABILIDAD DE LA DESINFORMACIÓN A TRAVÉS DE CHATBOTS

RESUMEN: El documento analiza los efectos adversos de la inteligencia artificial, particularmente en chatbots, enfocándose en los riesgos de la desinformación, la polarización y la discriminación algorítmica. Se explora la evolución de los chatbots, desde sistemas basados en reglas hasta modelos de lenguaje avanzados como A.L.I.C.E.®, Replika®, ChatGPT®, Bard® y DeepSeek®. Además, se propone la Inteligencia Artificial Explicada (XAI) como una solución para mitigar estos riesgos mediante la explicabilidad de los sistemas de IA. Finalmente, se evalúa la aplicación de los principios de XAI en diferentes chatbots, identificando sus fortalezas y debilidades en cuanto a explicabilidad, interpretabilidad, y ética. El texto concluye destacando la importancia de la XAI para un uso responsable y ético de la IA.

PALABRAS CLAVE: Inteligencia Artificial Explicable; Desinformación; Riesgos; Chatbots; Aprendizaje automático.

TABLE OF CONTENTS:

1. René Palacios Garita, "Artificial intelligence: challenges of explainability on disinformation through chatbots", *Latin American Journal of European Studies* 5, no. 2 (2025): 207 et seq.

2. Lawyer (BUAP), Economist (UIA), Master of Economic Analysis of Law (UCM-IUIOG), Master in Public Administration (INAP), PhD in Political Science, Administration and International Relations (UCM) and PhD in Law (UNAM). Professor at UNAM. E-mail: rpalaciosg@derecho.unam.mx. https://orcid.org/0000-0001-7658-0441.

lainability of machine learning chatbots; 8. Chatbot Cases: From A.L.I.C.E. to Deepseek; 9. Artificial Intelligence Explained (XAI) of the chatbot cases addressed; Final Considerations; References.

INTRODUCTION

This paper analyses the challenges of explainability of artificial intelligence (AI) in the information age. It explores how AI can exacerbate societal problems, such as polarization and the spread of misinformation. Algorithms can perpetuate existing biases, resulting in discriminatory decisions. It also examines how AI can be used to create disinformation and manipulation, undermining trust in institutions.

A central theme is the analysis of the risks associated with chatbots. These systems can generate fake news and fraudulent content. In the face of these challenges, the document highlights the need for Explainable Artificial Intelligence (XAI). The principles of XAI are discussed, such as explainability, interpretability, comprehensibility, understandability, transparency, legality and ethics. XAI seeks to improve the understanding of how AI systems make decisions.

The types of chatbots are also classified, from rule-based to machine learning. The evolution of chatbots is analyzed, from A.L.I.C.E.® to advanced models such as Replika®, ChatGPT®, Bard®, and DeepSeek®, comparing their capabilities in relation to XAI. Accordingly, how XAI applies to machine learning chatbots is explored.

Accordingly, the methodology used combines the responsible and ethical use of AI, driven by XAI, that essential to ensure that technology benefits society. Furthermore, Explained Artificial Intelligence (XAI) is proposed as a solution to mitigate these risks through explainability of AI systems.

The document emphasizes the importance of adopting an ethical and responsible approach to AI development, with special attention to XAI, to face the challenges of explainability and ensure that technology benefits society. It is therefore necessary to begin by addressing a general context that allows us to cognitively locate ourselves in this approach.

1. A BRIEF GENERAL CONTEXT

Artificial intelligence (AI), defined as "machine or software intelligence",[3] has experienced exponential growth in recent decades, transforming various aspects of daily life and generating a significant impact on society. This impact is manifested in multiple sectors: from health and recruitment to manufacturing and finance[4] with increasingly frequent interactions between humans and intelligent systems through text, voice and images.

This proliferation of AI has led to the development of various applications such as chatbots and virtual assistants.[5] However, the increasing adoption of AI also raises ethical and social concerns. One of the main fears is the potential of AI to exacerbate problems such as social polarization and questions about responsibility and accountability in the use of AI.[6]

In this sense, the impact of chatbots on the polarization of public opinion and the dissemination of misinformation generates biases in decision-making. This document takes into account examples of the use of AI to face challenges of explainability, which is why the capabilities of DeepSeek® and other advanced models in text generation are analyzed, as well as their potential to create content, which in the best-case scenario are expected to achieve positive results.[7]

These AI systems can process vast amounts of data at unmatched speed and accuracy, enabling businesses to streamline their operations and reduce costs. In the manufacturing sector, AI can monitor production, predict equipment failures, and improve product quality.[8] Even in the service field, chatbots and virtual assistants improve customer service and reduce the workload of human staff.

3. Svetlana Bialkova, *The Rise of AI User Applications: Chatbots Integration Foundations and Trends* (Cham: Springer, 2024), 3.
4. Jovan, Pehcevski, ed., *Chatbots and Text Generation* (Ashland: Arcler Press, 2024), 165.
5. James, Crowder, *AI Chatbots: The Good, the Bad, and the Ugly* (Cham: Springer Nature Switzerland, 2024), 2.
6. Mitisha, Gaur, "An In-Depth Exploration of Predictive Justice with AI," in *AI and Emerging Technologies: Automated Decision Making, Digital Forensics and Ethical Considerations*, ed. P. Pokhariyal, A. Patel, and S. Pandey, 1st ed. (Boca Raton: CRC Press, 2025), 241.
7. James, Crowder, *AI Chatbots: The Good, the Bad, and the Ugly*, 128.
8. Alkadash, Tamer, et al., "Maximizing Organizational Efficiency Through HR Information Systems: A Focus on Decision-Making in Tech Firms," in *AI and Business, and Innovation Research: Un-*

However, along with these benefits and opportunities, there are risks and adverse effects that cannot be ignored. One of the biggest fears is the potential for AI to exacerbate social and economic inequalities. AI algorithms can perpetuate existing biases in data, which can result in discriminatory decisions in areas such as hiring, access to credit, and criminal justice.[9]

In this context, one of the most critical aspects is risk management. AI systems can fail, be biased, or be used for malicious purposes, which could cause considerable harm. For example, facial recognition systems can be inaccurate and discriminate against certain groups of people and perpetuate existing biases in training data, leading to unfair outcomes. It is therefore important to recognize the recent evolution of risk that we are experiencing with this emerging technology.

2. RISK ASSOCIATED WITH AI

In general terms, risk can be understood as "the probability of an adverse event occurring, combined with the severity of its impact".[10] In the context of AI, this definition takes on more complex nuances. AI systems can generate risks through the loss of confidentiality, integrity or availability (CIA) of information and systems, thereby implying loss with adverse consequences on organizational operations, assets and people.[11]

Regarding AI, it can be classified as Weak or Narrow AI, designed to perform specific tasks. Most current AI applications, including chatbots,[12] Strong or General AI, with the ability to perform a variety of tasks as well as a human,[13]

---

derstanding the Potential and Risks of AI for Modern Enterprises, ed. B. Alareeni and I. Elgedawy, 1st ed. (Cham: Springer, 2024), 435.

9. Siva, Sindiramutty, et al., "Risk Assessment and Mitigation With Generative AI Models", in *Utilizing Generative AI for Cyber Defense Strategies,* ed. N. Jhanjhi (Hershey: IGI Global, 2025), 56.

10. Patrick, Hall, et al., *Machine Learning for High-Risk Applications: Approaches to Responsible AI*, First edition (Sebastopol: O'Reilly Media, 2023), 7.

11. Izzat, Alsmadi, "IT Risk and Security Management", in *The NICE Cyber Security Framework: Cyber Security Management,* ed. Alsmadi, I., et al., (Cham: Springer, 2020), 55.

12. Robert, Ciesla, *The Book of Chatbots: From ELIZA to ChatGPT* (Cham: Springer, 2024), 41.

13. Crowder, *AI Chatbots*, 101.

and Artificial Superintelligence,[14] which is around a hypothetical level of AI that would surpass human intelligence in all aspects.

In this context, in the information age, characterized by the proliferation of digital technologies and increasing global interconnection, the nature of risk has been transformed in all areas of society. AI is a distinctive feature of the information age, driving the development of systems that can learn, reason and make decisions autonomously.[15] In this, the integration of technologies such as AI with blockchain technology and quantum computing unlocks the potential for innovation and transformation in various industries.[16]

Under these circumstances, the integration of AI in various sectors has brought with it a series of benefits, but it has also exposed systems to new operational and security risks. In this regard, the risk of vulnerability of AI systems to cyber-attacks has become a central concern, since these attacks can compromise the integrity, confidentiality and availability of information and services.[17]

Therefore, AI offers tools to improve cybersecurity, can also be used by cybercriminals to create more sophisticated attacks[18] such as the proliferation of disinformation and the manipulation of information. This disinformation can undermine trust in institutions, polarize society and destabilize democratic processes. Hence, AI plays a role in the creation and dissemination of false content,[19] which increases the difficulty in distinguishing between reality and fiction

Accordingly, social media algorithms, which are designed to maximize user engagement, often amplify the spread of polarizing content and misinformation,

14. Bialkova, *Rise of AI User Applications*, 216.
15. Patrick, Murphy, "A Smart and Secure Healthcare System: Automated Methods for Diagnostics", in *The Emerging Role of AI-Based Expert Systems in Cyber Defense and Security*, ed. K. R. Bhatele et al., 1st ed. (New York: Nova Science Publishers, 2024), 305.
16. Petar, Radanliev, *The Rise of AI Agents: Integrating AI, Blockchain Technologies, and Quantum Computing*, First edition (Boston: Addison-Wesley, 2025), 44.
17. Jhanjhi, *Utilizing Generative AI*, 288.
18. Ravindra, Das, *Generative AI Phishing and Cybersecurity Metrics* (Boca Raton: Taylor & Francis Group, 2024), 45.
19. I., Almeida, *Responsible AI in the Age of Generative Models: Governance, Ethics and Risk Management*, 2024 edition (s.l.: Now Next Later AI, 2024), 9.

creating an environment where information is decontextualized.[20] Those, according to Rubin the proliferation of fake news and misinformation are therefore significant risks in the digital landscape, which has eroded trust in traditional media and made it difficult to distinguish what is fake.

In turn, the increasing automation of security tasks with AI also brings with it new risks. Attackers may attempt to fool intelligence tools by creating strategies designed to thwart authentication or threat detection.[21] To mitigate these risks, it is crucial that AI models used in security are robust and resistant to adversarial attacks, and that they are used in combination with human supervision.[22]

Specifically, risks associated with AI such as algorithmic biases, lack of transparency, and difficulty in understanding how AI systems make decisions can lead to unfair and discriminatory outcomes.[23] In this scenario, traditional risks persist, but new challenges related to technology, information, and global interconnection have been added, which introduce an additional layer of complexity with these types of risks.[24]

A red flag is certainly found in the context of autonomous weapons. Especially since this lack of transparency is even more worrying, as it can lead to situations where humans cannot control or interrogate the machine's decisions. As mentioned by Buchanan & Imbrie, one critic of autonomous weapons points out, human control is reduced to pressing an "I think" button without the possibility of understanding the underlying logic.[25]

One of the main risks is therefore the danger of relying too much on AI capabilities without understanding its limitations.[26] AI models, although advanced, are

20. Victoria, Rubin, *Misinformation and Disinformation: Detecting Fakes with the Eye and AI* (Cham: Springer International Publishing, 2022), 80.
21. Jhanjhi, *Utilizing Generative AI*, 181.
22. Théo, Lepage-Richer, "Adversariality in Machine Learning Systems: On Neural Networks and the Limits of Knowledge," in *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, ed. J. Roberge and M. Castelle, 1st ed. (Cham: Palgrave Macmillan, 2021), 216.
23. Simon, Lindgren, *Critical Theory of AI* (Cambridge: Polity Press, 2024), 163.
24. Pieter, Verdegem, *AI for Everyone? Critical Perspectives* (London: University of Westminster Press, 2021), 241.
25. Ben, Buchanan and Andrew Imbrie, *The New Fire: War, Peace, and Democracy in the Age of AI* (Cambridge, MA: MIT Press, 2022), 149.
26. Crowder, *AI Chatbots*, 218.

not infallible. They can be vulnerable to errors, biases, and adversarial attacks.[27] Blind trust in AI can lead to excessive delegation of critical tasks without proper human oversight, which could result in negative consequences.[28] Therefore, people may be less tolerant of technology errors than of human errors and may trust the results of systems less than the opinions of other people.[29] In accordance with the risks addressed, it is important to recognize the evolution that this has had over recent times. Accordingly, the following section is prepared to address the evolution that it has had.

3. CHALLENGES OF EXPLAINABILITY OF AI

Challenges of AI refer to negative or harmful consequences arising from the design, development, implementation, or use of AI systems.[30] Accordingly, AI can be used to personalize disinformation, tailoring messages to each user's specific beliefs and biases. This personalization makes disinformation even more persuasive and difficult to refute, as it is designed to resonate with the person's preconceptions. The combination of AI-generated content and message personalization creates an environment in which fake news thrives[31]. However, according to Ciesla it is important to note that disinformation is not always intentional as many chatbots learn from data available on the internet, which may contain biased or erroneous information and AI's tendency to "hallucinate" and fill in gaps with fabricated information can generate this type of disinformation.

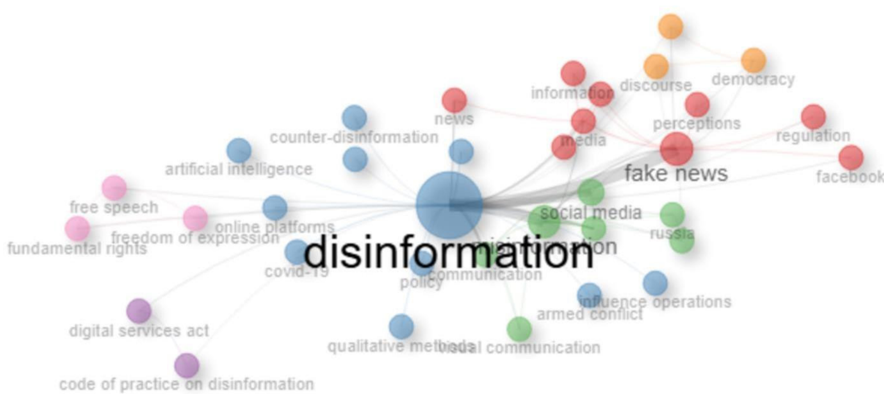
The speed at which AI-powered misinformation spreads make any timely correction or denial extremely difficult, and it often comes too late. Furthermore, the nature of AI increases the difficulty of establishing accountability, as it increases the lack of traceability and the anonymity of authorship. Disinformation

27. Lepage-Richer, "Adversariality in Machine Learning Systems," 201..
28. Gaur, "In-Depth Exploration of Predictive Justice," 266..
29. Andreas, Moring, *AI on the Job: Guide to Successful Human-Machine Collaboration* (Cham: Springer, 2022), 124.
30. Bialkova, *Rise of AI User Applications*, 153.
31. Ciesla, *Book of Chatbots*, 127.

is a direct attack on the collective face of freedom of expression, undermining a society's right to the truth and the free and timely exercise of political rights.

This is why we need to understand the difference between disinformation and the manipulation of information conceived as "malinformation"[32] that lies in the intention behind it. For example, deepfake technology allows for the creation of fake videos and audios that can be used to defame individuals or manipulate political events.[33] These deepfakes are so convincing that they can make people stop believing what they see. The impact of this manipulation is enormous and raises serious questions about the authenticity of information circulating online.[34]Ultimately, both unintentional and intentional disinformation, despite not having the same motivation, can have equally harmful effects by distorting reality and confusing people. We can observe the different conceptual derivations according to the following graph that interrelates the contextual elements addressed:

Figure 1. Conceptual considerations around disinformation



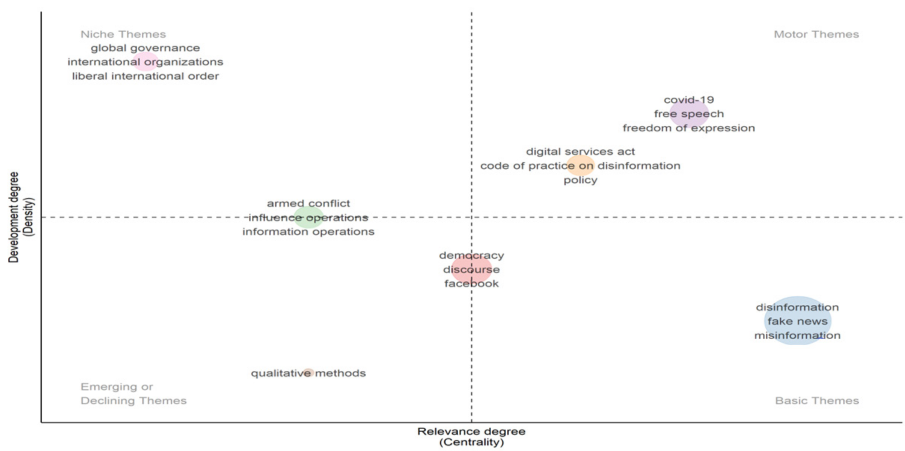
Source: Prepared according to what was proposed by Massimo Aria and Corrado Cuccurullo, "Bibliometrix: An R-Tool for Comprehensive Science Mapping Analysis," *Journal of Informetrics* 11, no. 4 (2017): 959-975.

---

32. "Misinformation, Disinformation, and Malinformation: A Guide," Princeton Public Library, last modified February 7, 2025, https://princetonlibrary.org/guides/misinformation-disinformation-malinformation-a-guide/
33. Divyansh, Shukla and Anshul, Pandey, "An Invisible Threat to the Security of Nations in the Age of 'Deepfakes'," in *AI and Emerging Technologies: Automated Decision Making, Digital Forensics and Ethical Considerations*, ed. P. Pokhariyal, A. Patel, and S. Pandey, 1st ed. (Boca Raton: CRC Press, 2025), 375.
34. Buchanan and Imbrie, *New Fire*, 199.

One of the main challenges is that the datasets used to train AI often reflect existing prejudices and inequalities in society.[35] If the training data contains racial, gender or socioeconomic biases, AI models will learn and replicate these biases, even if they are not explicitly programmed to do so.[36] For example, according to Verdegem, if a facial recognition system is trained primarily on images of people of a specific ethnicity, it is likely to perform poorly when identifying people of other ethnicities.[37] However, their growing presence raises questions about the transparency and explainability of their decisions and responses. Understanding how these systems process information, and especially how they can spread misinformation or unintentional biases has become central. To exemplify this, the following thematic map was generated:

Figure 2. Thematic map around challenge of explainability and disinformation



Source: Prepared according to what was proposed by Massimo Aria and Corrado Cuccurullo, "Bibliometrix: ..." 959-975.

Basic topics, such as disinformation, fake news, and misinformation, are at the heart of the debate on the reliability of AI systems and the need for explainable mechanisms that guarantee traceability and veracity in the responses generated. Key topics, such as policy, freedom of expression, and COVID-19, reflect established areas that articulate the discussion between regulation, freedom

35. Lepage-Richer, "Adversariality in Machine Learning Systems," 219.
36. Lindgren, *Critical Theory of AI*, 144.
37. Verdegem, *AI for Everyone?*, 303.

of expression, and digital governance. Specialized topics, such as global governance and international organizations, address the political and ethical dimensions of global AI regulation, while emerging topics, such as qualitative methods, highlight methodological approaches that seek to understand human-computer interaction and the perception of transparency.

In addition to biases in data, the way algorithms are designed and implemented can also contribute to discrimination. For instance, algorithms can be opaque and difficult to understand, making it difficult to identify and correct biases. Thus, lack of transparency in algorithms and the AI decision-making process can make it even harder to detect and correct discrimination.[38]

4. CHALLENGES OF EXPLAINABILITY OF CHATBOTS

Chatbots, defined primarily as "software designed to interact with people through text or voice".[39] Therefore, it is essentially a computer program created to simulate human conversations.[40] They are located within Weak or Narrow AI because their function is to perform specific tasks such as answering questions, assisting in purchases, or maintaining conversations, which use natural language processing (NLP) to understand the questions and generate responses.

The problem arises when the information provided by these systems reflects bias, falsehood or opinions without considering the multiplicity of impacts they have on complex issues. Therefore, chatbots and other AI-based systems can produce fake news, fraudulent research articles and social media posts that are difficult to distinguish from truthful information, and this type of content can appear highly authentic, making it difficult to detect and verify.[41]

The ease with which this content can be produced and disseminated therefore allows disinformation to spread rapidly, reaching large numbers of people in

38. Lindgren, *Critical Theory of AI*, 142.
39. Crowder, *AI Chatbots*, 22.
40. Pehcevski, *Chatbots and Text Generation*, 32.
41. Ciesla, *Book of Chatbots*, 127..

a short period of time.[42] The case of Microsoft's "Tay", which became offensive within 24 hours, demonstrates how chatbots can learn inappropriate behavior from human interactions online.

Crowder give us another example. It is the case of the chatbot "Tessa",[43] which was quickly removed after providing potentially harmful information and inappropriate advice, illustrating the dangers of replacing human professionals with automated systems. Therefore, the lack of real empathy and the inability to understand the subtleties of human language can lead to negative outcomes. It is therefore necessary to generate strategies to enable the reduction of these challenges of explainability and in light of this, XAI emerges as a possibility.

5. EXPLAINED ARTIFICIAL INTELLIGENCE (XAI) FOR THE CHALLENGES OF EXPLAINABILITY OF AI

The need for transparent explanations in AI decision-making, or better known as XAI, is most relevant in high-risk situations and important decisions. The complexity of some systems and the subjectivity inherent in human decision-making can make it difficult to create complete and meaningful explanations.[44] This results in a tension between transparency and other objectives such as privacy and intellectual property.[45]

Explainability is a fundamental pillar of responsible AI, along with fairness, privacy, security, robustness, and accountability. UNESCO, for its part, promotes transparency and social benefit orientation as basic principles. Accordingly, "Explainability involves making the results of AI systems intelligible and providing information about them".[46] For explainability to be effective, it is necessary to go beyond diffuse concepts. The practical implementation of explainability requires ability to provide explanations for specific information.

42. Rubin, *Misinformation and Disinformation*, 266.
43. Crowder, *AI Chatbots*, 98.
44. Phillips-Wren, Gloria, Nikhil Ichalkaranje, and Lakhmi Jain, *Intelligent Decision Making: An AI-Based Approach* (Berlin: Springer, 2008), 10.
45. Almeida, *Responsible AI*, 172-173.
46. UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (Paris: UNESCO, 2021), 23.

For that reason, one effect of a lack of explainability, then, is to make it difficult to identify and correct errors in AI models. If we do not understand how a system works, it is difficult to diagnose and fix any problems that may arise. For example, a system that incorrectly predicts disease risks can lead to poor healthcare decisions, but without proper transparency, errors may not be detected until it is too late.[47]

The explainability is essential because, without it, there is a risk that algorithmic decisions will be based on incomplete or biased data, generating negative consequences for the efficiency and equity of public policies Therefore ECLAC refers that the lack of transparency and explainability, among others, in the decision-making of AI raises problems that undermine the trust and acceptance of this technology.[48] If we cannot understand how a decision was made, it is difficult to determine who is responsible for any resulting error or damage. Given this, Bialkova proposes a Taxonomy to be able to address AI from a responsible approach, combining it with the principles related to XAI, referring to the following:

Table 1. Principles of Explained Artificial Intelligence (XAI)

Principle	**Description**
Explainability	The ability of the system to explain its reasoning and decisions to users.
Interpretability	The extent to which an observer can understand the causes of a decision.
Comprehensibility	The ease with which users can perceive how the system works.
Understandability	The ease with which users can understand how the system works.
Transparency	The ability of the system to ensure that XAI objectives are met by revealing its operations, data and algorithms.
Legality	The ability of the system to ensure that XAI objectives are met with applicable laws and regulations

47. Gaur, "In-Depth Exploration of Predictive Justice," 328-329.
48. ECLAC, *Latin American Artificial Intelligence Index (ILIA) 2025* (Santiago, Chile: CEPAL, 2025), 133.

Ethics	The system's ability to ensure that XAI's objectives are based on ethical standards that ensure responsible use.

Source: Prepared according to what was proposed by Bialkova, S. *The Rise of AI User Applications: Chatbots Integration Foundations and Trends*. Springer, 2024.198.

Accordingly, efforts to create more transparent and explainable AI systems are crucial to foster trust in this technology and ensuring its ethical and responsible use. The challenge lies not only in improving technical tools, but also in considering the human and social aspects of interactions with AI. In line with this proposal, we address the case of the announced chatbots.

6. CHATBOT TYPOLOGY

Regarding the classification of chatbots, they are divided into several categories according to their functionality, such as menu-based chatbots (those that offer predefined options for users to select, such as fast-food kiosks, automated customer service telephone systems), voice-based digital assistant chatbots such as Siri or Alexa that are designed to perform specific tasks and answer questions. In turn, we have rule-based chatbots that work by following a predefined set of rules and responses,[49] those based on cognition by simulating human cognitive processes[50] and those based on machine learning and neural networks.[51]

7. ARTIFICIAL INTELLIGENCE EXPLAINED (XAI) TO FACE THE CHALLENGES OF EXPLAINABILITY OF MACHINE LEARNING CHATBOTS

Based on these characteristics, we can propose this typology from the scheme proposed by Bailkova, as well as the risk approach that has been addressed at the moment, in this case applied to machine learning chatbots and neural networks. In this regard, it can be stratified:

49. Pehcevski, *Chatbots and Text Generation*, 60.
50. Crowder, *AI Chatbots*, 30.
51. Bialkova, *Rise of AI User Applications*, 218.

Table 2. Principles of Explained Artificial Intelligence (XAI) in machine learning chatbots

Relationship	Application of principle	Risk
EXPLAINABILITY		
Uses machine learning algorithms to understand context and learn from past interactions	Minimizing those models can be "black boxes" that are difficult to interpret, although XAI techniques seek to improve explainability.	Possibility of biases in training data, "hallucinations" or incorrect responses.
INTERPRETABILITY		
They learn from past interactions and data	It depends on the architecture of the model; deeper models (deep learning) are black boxes that are difficult to interpret.	Incorrect or meaningless answers if training data is insufficient or biased.
COMPREHENSIBILITY		
It uses natural language processing (NLP) to	It depends on the quality of the model and the training	Over training and data, results can be unpredictable
understand the context of the question and generate answers.	data. It can be high if the model is well trained, but it may fail on complex or ambiguous questions.	(emergent behavior), potential use for misinformation.
UNDERSTANDABILITY		
Machine learning to understand user context and language, learning from each interaction.	Learn from every interaction, improving your ability to understand and respond over time.	Systems that depend on large amounts of data for training information may not be diverse due to contaminated information.
TRANSPARENCY		
They seek to imitate human conversation and learn from each interaction.	Significant advancement in the ability of chatbots to understand and respond to natural language.	The decision-making logic is based on learning algorithms that can be difficult to understand even for developers.
LEGALITY		
Compliance with AI laws in development.	Responsibility for the quality and accuracy of information. Must avoid the dissemination of false information and ensure the privacy of user data.	Biases in training data leading to discriminatory or unfair responses. Difficulty in controlling implicit learning and its emergent behavior when generating false or unverified information.

ETHICS		
Consideration of how these systems can contribute to the common good and avoid exacerbating social divisions	Information for the user who should be aware that the chatbot may not be perfect and whoever trains the chatbot should ensure quality	Information asymmetries on complex issues that ultimately depend on the quality of the data.
	and avoid bias, as well as the misuse of captured personal data.	

Source: Prepared according to what was proposed by Bialkova, *The Rise of AI User Applications*:..." 198.

8. CHATBOT CASES: FROM A.L.I.C.E. TO DEEPSEEK

The evolution of chatbots began from simple rule-based programs to today's sophisticated language models. For example, A.L.I.C.E.® (Artificial Linguistic Internet Computer Entity) was the Pioneer of Pattern-Based Conversation. A.L.I.C.E.®, created by Dr. Richard Wallace in 1995, represents an earlier era in the development of chatbots.[52] Inspired by ELIZA, A.L.I.C.E.® uses "Artificial Intelligence Markup Language" (AIML), a language specifically designed to structure conversations, which works by defining input patterns that are associated with predefined responses, similar to a template system.[53] In addition, Pehcevski argue that unlike current language models, A.L.I.C.E.® does not learn from interactions dynamically, but rather relies on pre-established patterns, which limits its ability to handle complex or ambiguous conversations.[54]

In stark contrast to A.L.I.C.E.® 's rules-based approach, Replika®, launched in 2017, focuses on creating a virtual companion with whom users can establish an emotional connection, using natural language processing (NLP), machine learning, and artificial neural networks (ANN) to understand the user's emotions and preferences, personalizing its responses accordingly.[55] Therefore, Replica's primary goal is not to answer questions or complete tasks, but to provide a safe space for users to express their thoughts and feelings.

52. Crowder, *AI Chatbots*, 8.
53. Ciesla, *Book of Chatbots*, 53.
54. Pehcevski, *Chatbots and Text Generation*, 32.
55. Crowder, *AI Chatbots*, 57.

For its part, ChatGPT®, specialized in text generation and launched in 2022 by OpenAI, represents a quantum leap in the capabilities of chatbots. Based on the large language models (LLM) GPT-3.5 and GPT-4, ChatGPT® can generate text similar to how a human would do it, which unlike A.L.I.C.E.® , which is based on pre-established patterns, and Replika®, whose focus is emotional interaction, ChatGPT® learns from huge amounts of text to generate coherent and relevant responses.[56]

Bard®, announced by Google in 2023, emerges as a direct competitor to ChatGPT®, utilizing the power of the LaMDA and PaLM 2 LLMs. Like ChatGPT®, Bard® is capable of generating high-quality text, but a crucial difference lies in its ability to search the internet for information in real-time. This capability allows it to provide more up-to-date and contextualized responses. Furthermore, Bard® integrates with other Google services and third-party tools such as Adobe and Spotify, making it easy to use in a variety of tasks and workflows.[57] An example of this would be asking Bard® for information about the latest news or the weather forecast, something that ChatGPT®, which does not have access to the internet in real-time, would not be able to do.

Finally, DeepSeek® emerges as a specialized model, particularly excelling at coding and math tasks, outperforming other models on specific benchmarks, generating code in multiple languages, solving math problems, and explaining algorithms. Unlike ChatGPT® and Bard®, DeepSeek® focuses on areas and its multilingual capability, i.e., it can process multilingual data on an extraordinary scale, including not only widely spoken languages such as English and Mandarin, but also regional dialects and minority languages.[58]

56. Ciesla, *Book of Chatbots*, 66-67.
57. Ciesla, *Book of Chatbots*, 77.
58. Alexander, Harrington, *The DeepSeek Disruption: How China's Groundbreaking AI Model Is Transforming the Global Tech Race* (Uxbridge, UK: Ria Christie Collections, 2025), 19.

9. ARTIFICIAL INTELLIGENCE EXPLAINED (XAI)OF THE CHATBOT CASES ADDRESSED

Explainable Artificial Intelligence (XAI) has become an essential field in the development of chatbots, seeking to make the decision-making processes of the models transparent. The implementation of XAI in chatbots entails numerous challenges since models such as LLMs are extremely powerful, but their complexity makes it difficult to explain their behavior. According to Bialkova's proposal (2024), an analysis is made regarding each of these mentioned chatbots, in order to make it cognitive to understand the XAI model that can be discerned from them:

Table 3. Principles of Explained Artificial Intelligence (XAI) in A.L.I.C.E. ®, Replika®, ChatGPT®, Bard® and DeepSeek® machine learning chatbots **A.L.I.C.E.®**	Replika®	ChatGPT®	Bard®	DeepSeek®
EXPLANABILITY				
It uses predefined rules and language patterns, making it easy to understand how it works, but the complexity of AIML can be difficult to track.	It uses machine learning models, but its internal logic is opaque.	Large language model (LLM) that operates as a "black box", with complex patterns that are difficult to track.	Like ChatGPT®, based on complex deep learning models. Its internal logic is difficult to interpret.	Decision making is complex, and reasoning processes are not transparent.
INTERPRETABILITY				
It uses if-then rules, which makes it easy to follow, but the number of rules can make it difficult to maintain.	Although based on ML, it is oriented towards simpler and more affective interactions than complex information tasks.	Complex, black box-like architecture, difficulty in understanding why it gives a specific response.	Like ChatGPT® in its black box nature; opaque internal decision processes.	Like other LLMs, it is characterized by the complexity of its models and the difficulty of tracing its decision processes.
COMPREHENSIBILITY				
It uses flexible pattern matching rules, it may look human, but its understanding is limited.	He tries to be an "online friend," so intelligibility focuses on simulated empathy rathert than understanding.	It uses large language models and can generate coherent and contextually relevant text, but can fail at complex reasoning.	Has the ability to search for updated information on the Internet, which improves understanding and response capacity.	Possibility of errors in coding tasks, problems with using real- world information.

UNDERSTANDABILITY				
If the question is more complex or outside of your range of patterns, you may not respond adequately.	If the user Expects deep emotional understanding, they might be disappointed.	A user might ask to write an essay, receiving a well-written text but without understanding whether the information is completely accurate.	Like ChatGPT®, it can generate misinformation and responses can be inconsistent.	It can be opaque in its internal workings, making it difficult to understand and verify.
TRANSPARENCY				
Although its code is available as open source, its internal logic based on complex rules (AIML) can be difficult to understand for non-technical users.	It is not clear how the emotional connections he advertises are formed.	It uses deep learning models, whose internal logic is difficult to interpret.	Like ChatGPT®, based on machine learning models with internal logic that is difficult to understand.	Shares the risks ofChatGPT® and Bard®: biases, generation of incorrect or false information, and unpredictable behavior.
LEGALITY				
No specific cases of legal problems are mentioned.	If it is perceived as a replacement for professional support, there could be legal risks.	It raises issues about the veracity and reliability of information and cannot do causal analysis or research.	No specific legal cases regarding Bard® are discussed, but it is mentioned that it competes with ChatGPT®, so similar risks are assumed.	They do not detail specific legal cases for DeepSeek®, but being an advanced LLM- based AI technology, it is assumed that it shares the risks of similar models.
ETHICS				
Open source, which allows for community contribution but also means there is less centralized control over its use and potential misuse.	Relationship with "digital privacy", which points to the specific thical risk.	Discussion on deepfakes, which could be associated with ethical risks.	No specific information on ethical concerns or unique risks but has parallels with ChatGPT®.	The fact that it is an open-source mode implies greater distribution, which could mean greater potential for misuse.

ource: Prepared according to the referenced papers, as well as what was proposed by Bialkova, *Rise of AI User Applications*:…", 190.

As observed regarding XAI, it is essential for all chatbots, addressing the need for understanding, trust, and accountability in AI systems. Each chatbot presents its own challenges and opportunities for XAI, from the transparency of rules in A.L.I.C.E.® to the complexity of language models in ChatGPT®, Bard®, and DeepSeek®. XAI not only improves the quality of chatbots but also ensures that their impact on society is positive and equitable. XAI also requires evaluating the suitability of the data used by the models.

With this, it can be expected that XAI seeks to minimize biases and ensure fairness, which implies interactivity, allowing users to question the behavior of a chatbot and influence the responses and outcomes, especially in current times when the risks associated with AI, and specifically the risks associated with chatbots, continue to be an area of opportunity to be solved.

FINAL CONSIDERATIONS

While AI offers benefits across several sectors, it also poses significant risks such as social polarization, misinformation, cyberattacks and perpetuation of bias. It is crucial to recognize both the potential and dangers of this technology.

Within the AI modalities, chatbots present risks as they can produce fake news and fraudulent content. Lack of empathy and inability to understand the subtleties of language can lead to negative outcomes. Therefore, the need for Explainable Artificial Intelligence (XAI) is highlighted to foster trust and ensure ethical use of AI based on principles such as explainability, interpretability, intelligibility, comprehensibility, transparency, legality and ethics.

XAI is essential for all chatbots, seeking understanding, trust, and accountability. XAI not only improves the quality of chatbots, but ensures positive social impact, minimizing bias and allowing users to question chatbot behavior. The risks of AI, and specifically of chatbots, remain an area of opportunity to be resolved.

Chatbots, despite their usefulness, can spread misinformation and display inappropriate behavior, highlighting the need for human oversight and clear ethical principles in their development. The paper emphasizes the importance of a responsible and ethical approach in AI development, with a special focus on XAI, to face challenges of explainability and ensure that technology benefits society.

In conclusion, the ethical and responsible development and implementation of Artificial Intelligence (AI) require a holistic and multidisciplinary approach that encompasses technical, social, legal and ethical aspects. The rapid proliferation of AI, especially in the form of chatbots and machine learning systems, has brought substantial benefits, but has also brought with it significant risks that cannot be ignored. It is crucial, therefore, to adopt measures that ensure that AI is used for the common good and does not exacerbate inequalities or undermine trust in information.

The future of AI therefore demands responsibility: it concludes with a call to action to adopt a responsible approach to AI development, ensuring that technology benefits society as a whole and does not exacerbate existing inequalities or risks. XAI emerges as a fundamental pillar for building a future in which AI is used in an ethical, equitable and responsible manner.

REFERENCES

Almeida, I. *Responsible AI in the Age of Generative Models: Governance, Ethics and Risk Management*. 2024 edition. s.l.: Now Next Later AI, 2024.

Alsmadi, Izzat, "IT Risk and Security Management". In *The NICE Cyber Security Framework: Cyber Security Management*, ed. I. Alsmadi, C. Easttom and L. Tawalbeh, 1st ed., 55–78. (Cham: Springer, 2020).

Bialkova, Svetlana. *The Rise of AI User Applications: Chatbots Integration Foundations and Trends*. Cham: Springer, 2024.

Buchanan, Ben and Andrew Imbrie. *The New Fire: War, Peace, and Democracy in the Age of AI*. Cambridge, MA: MIT Press, 2022.

Ciesla, Robert. *The Book of Chatbots: From ELIZA to ChatGPT*. Cham: Springer, 2024.

Crowder, James. *AI Chatbots: The Good, the Bad, and the Ugly*. Cham: Springer Nature Switzerland, 2024.

Das, Ravindra. *Generative AI Phishing and Cybersecurity Metrics*. Boca Raton: Taylor & Francis Group, 2024.

ECLAC. *Latin American Artificial Intelligence Index (ILIA) 2025*. Santiago, Chile: CEPAL, 2025.

Gaur, Mitisha. "An In-Depth Exploration of Predictive Justice with AI." In *AI and Emerging Technologies: Automated Decision Making, Digital Forensics and Ethical Considerations*, edited by P. Pokhariyal, A. Patel, and S. Pandey, 241–334. 1st ed. Boca Raton: CRC Press, 2025.

Hall, Patrick, James Curtis and Parul Pandey. *Machine Learning for High-Risk Applications: Approaches to Responsible AI*. First edition. Sebastopol: O'Reilly Media, 2023.

Harrington, Alexander. *The DeepSeek Disruption: How China's Groundbreaking AI Model Is Transforming the Global Tech Race*. Uxbridge, UK: Ria Christie Collections, 2025.

Lepage-Richer, Théo. "Adversariality in Machine Learning Systems: On Neural Networks and the Limits of Knowledge." In *The Cultural Life of Machine Learning: An Incursion into Critical AI Studies*, edited by J. Roberge and M. Castelle, 199–224. 1st ed. Cham: Palgrave Macmillan, 2021.

Lindgren, Simon. *Critical Theory of AI*. Cambridge: Polity Press, 2024.

Moring, Andreas. *AI on the Job: Guide to Successful Human-Machine Collaboration*. Cham: Springer, 2022.

Murphy, Patrick. "A Smart and Secure Healthcare System: Automated Methods for Diagnostics." In *The Emerging Role of AI-Based Expert Systems in Cyber Defense and Security*, edited by K. R. Bhatele et al., 1st ed., 305–332. New York: Nova Science Publishers, 2024.

Pehcevski, Jovan. ed. *Chatbots and Text Generation*. Ashland: Arcler Press, 2024.

Phillips-Wren, Gloria, Nikhil Ichalkaranje, and Lakhmi Jain. *Intelligent Decision Making: An AI-Based Approach*. Berlin: Springer, 2008.

Princeton Public Library. "*Misinformation, Disinformation, and Malinformation: A Guide*." Last modified February 7, 2025. https://princetonlibrary.org/guides/misinformation-disinformation-malinformation-a-guide/

Radanliev, Petar. *The Rise of AI Agents: Integrating AI, Blockchain Technologies, and Quantum Computing*. First edition. Boston: Addison-Wesley, 2025.

Rubin, Victoria. *Misinformation and Disinformation: Detecting Fakes with the Eye and AI*. Cham: Springer International Publishing, 2022.

Shukla, Divyansh, and Anshul Pandey. "An Invisible Threat to the Security of Nations in the Age of 'Deepfakes'." In *AI and Emerging Technologies: Automated Decision Making, Digital Forensics and Ethical Considerations*, edited by P. Pokhariyal, A. Patel, and S. Pandey, 1st ed., 204–226. Boca Raton: CRC Press, 2025.

Siva, Sindiramutty, Noor Zaman Jhanjhi, Rehan Akbar, Tariq Rahim Soomro and Mustansar Ali Ghazanfar. "Risk Assessment and Mitigation With Generative AI Models", in *Utilizing Generative AI for Cyber Defense Strategies*, ed. N. Jhanjhi (Hershey: IGI Global, 2025).

Tamer, Alkadash, Mohammed Allaymouni, Alreem Khalid Almuslemani, and Yaser Saleh Ebrahim. "Maximizing Organizational Efficiency Through HR Information Systems: A Focus on Decision-Making in Tech Firms." In *AI and Business, and Innovation Research: Understanding the Potential and Risks of AI for Modern Enterprises*, edited by B. Alareeni and I. Elgedawy, 1st ed., 435–458. Cham: Springer, 2024.

UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. Paris: UNESCO, 2021.

Verdegem, Pieter. *AI for Everyone? Critical Perspectives*. London: University of Westminster Press, 2021.

Received on 16/10/2025

Approved on 07/11/2025