ADVERSE EFFECTS OF AI. RISKS IN THE INFORMATION AGE, DISINFORMATION AND MACHINE LEARNING CHATBOTS WITH EXPLAINABLE AI (XIA) APPROACH: FROM A.L.I.C.E. TO DEEPSEEK

EFEITOS ADVERSOS DA IA. RISCOS NA ERA DA INFORMAÇÃO, DESINFORMAÇÃO E APRENDIZADO DE MÁQUINA CHATBOTS COM ABORDAGEM DE IA EXPLICÁVEL (XIA): DE A.L.I.C.E. A DEEPSEEK

DOI: 10.19135/revista.consinter.00021.09
Received/Recebido 12/02/2025 – Approved/Aprovado 21/05/2025

René Palacios Garita[1] *– https://orcid.org/0000-0001-7658-0441*

Abstract

The paper analyses the adverse effects 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®. This entails the hypothesis that the increasing sophistication of chatbots, from rule-based systems to advanced language models, leads to a rise in the risks of misinformation, polarization, and algorithmic discrimination, thus, Explainable Artificial Intelligence (XIA) emerges as a crucial tool to mitigate these adverse effects, although its application in different chatbots presents strengths and weaknesses in terms of explainability. The methodology used combines the responsible and ethical use of AI, driven by XIA, that essential to ensure that technology benefits society. Furthermore, Explained Artificial Intelligence (XIA) is proposed as a solution to mitigate these risks through explainability of AI systems. The results suggest a relationship in the application of XIA principles in different chatbots is evaluated, identifying their strengths and weaknesses in terms of explainability principles. The paper concludes by highlighting the importance of XAI for a responsible and ethical use of AI.

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

Resumo

O artigo analisa os efeitos adversos da inteligência artificial, particularmente em chatbots, com foco nos riscos de desinformação, polarização e discriminação

[1] 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 Political and Social Sciences (SEP). Student at the UCM and graduated at the UNAM. E-mail: renepala@ucm.es, https://orcid.org/0000-0001-7658-0441.

I hereby declare, for all appropriate purposes, that generative artificial intelligence was used in the preparation of this article in the sense that the object of study examined herein was the identification of the particular features of generative artificial intelligence, as described in the document. Confirmed/authorized, I include this information in the article.

algorítmica. A evolução dos chatbots é explorada, de sistemas baseados em regras a modelos de linguagem avançados como A.L.I.C.E.®, Replika®, ChatGPT®, Bard® e DeepSeek®. Isso envolve a hipótese de que a crescente sofisticação dos chatbots, de sistemas baseados em regras a modelos de linguagem avançados, leva a um aumento nos riscos de desinformação, polarização e discriminação algorítmica, portanto, a Inteligência Artificial Explicável (XIA) surge como uma ferramenta crucial para mitigar esses efeitos adversos, embora sua aplicação em diferentes chatbots apresente pontos fortes e fracos em termos de explicabilidade. A metodologia usada combina o uso responsável e ético da IA, impulsionado pela XIA, que é essencial para garantir que a tecnologia beneficie a sociedade. Além disso, a Inteligência Artificial Explicada (XIA) é proposta como uma solução para mitigar esses riscos por meio da explicabilidade dos sistemas de IA. Os resultados sugerem que uma relação na aplicação dos princípios XIA em diferentes chatbots é avaliada, identificando seus pontos fortes e fracos em termos de princípios de explicabilidade. O artigo conclui destacando a importância do XAI para um uso responsável e ético da IA.

Palavras-chave: Inteligência Artificial Explicável. Desinformação. Riscos. Chatbots. Aprendizado de máquina.

1 INTRODUCTION

This paper analyses the adverse effects 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.

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 (XIA). The principles of XIA are discussed, such as explainability, interpretability, comprehensibility, understandability, transparency, legality and ethics. XIA seeks to improve the understanding of how AI systems make decisions.

This entails the hypothesis that the increasing sophistication of chatbots, from rule-based systems to advanced language models, leads to a rise in the risks of misinformation, polarization, and algorithmic discrimination, thus, Explainable Artificial Intelligence (XIA) emerges as a crucial tool to mitigate these adverse effects, although its application in different chatbots presents strengths and weaknesses in terms of explainability.

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 XIA. Accordingly, how XIA applies to machine learning chatbots is explored.

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

Thus, the document emphasizes the importance of adopting an ethical and responsible approach to AI development, with special attention to XIA, to mitigate adverse effects and ensure that the technology benefits society. It is therefore

necessary to begin by addressing a general context that allows us to cognitively locate ourselves in this approach.

The results suggest a relationship in the application of XIA principles in different chatbots is evaluated, identifying their strengths and weaknesses in terms of explainability principles. The paper concludes by highlighting the importance of XAI for a responsible and ethical use of AI.

2 ARTICLE DEVELOPMENT

Artificial intelligence (AI), defined as "machine or software intelligence" (Bialkova, 2024:3), 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 (Pehcevski, 2024:165) 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 (Crowder, 2023:2). 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 (Gaur, 2025:241).

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 generate adverse effects, 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 (Crowder, 2023:128).

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 (Tamer M. et al., 2024:435). Even in the service field, chatbots and virtual assistants improve customer service and reduce the workload of human staff.

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. (Jhanjhi, 2025:212).

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 (Hall, et al., 2023:174). It is therefore important to recognize the recent evolution of risk that we are experiencing with this emerging technology.

In general terms, risk can be understood as "the probability of an adverse event occurring, combined with the severity of its impact" (Hall, et al., 2023:7). 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 (Alsmadi, et al., 2020:55).

Regarding AI, it can be classified as Weak or Narrow AI, designed to perform specific tasks. Most current AI applications, including chatbots (Ciesla, 2024:41), Strong or General AI, with the ability to perform a variety of tasks as well as a human (Crowder, 2023:101), and Artificial Superintelligence (Bialkova, 2024:216), 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. Artificial intelligence (AI) is a distinctive feature of the information age, driving the development of systems that can learn, reason and make decisions autonomously (Murphy, 2024:305). In this, the integration of technologies such as artificial intelligence with blockchain technology and quantum computing unlocks the potential for innovation and transformation in various industries (Randaliev, 2025:44).

Under these circumstances, the integration of Artificial Intelligence (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 (Jhanjhi, 2025:288).

Therefore, artificial intelligence (AI), although it offers tools to improve cybersecurity (Jhanjhi, 2025:30), can also be used by cybercriminals to create more sophisticated attacks (Das, 2025:45) 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 (Almeida, 2024:9)., 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, creating an environment where information is decontextualized (Rubin, 2022:80). Those, according to Rubin (2022) 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 its new risks. Attackers may attempt to fool intelligence tools by creating strategies designed to thwart authentication or threat detection (Jhanjhi, 2025:181). 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 (Lepage-Richer, 2020:216).

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 (Lindgren, 2023:163). 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 (Verdegem, 2021:241).

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 (2022), 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 (p.149).

One of the main risks is therefore the danger of relying too much on AI capabilities without understanding its limitations (Crowder, 2023:128). AI models, although advanced, are not infallible. They can be vulnerable to errors, biases, and adversarial attacks (Lepage-Richer, 2020:201). Blind trust in AI can lead to excessive delegation of critical tasks without proper human oversight, which could result in negative consequences (Gaur, 2025:266). 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 (Moring, 2022:124). 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.

Adverse effects of AI refer to negative or harmful consequences arising from the design, development, implementation, or use of AI systems (Bialkova, 2024:153). 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 (Ciesla, 2024:127). However, according to Ciesla (2024) 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 (p.128).

The difference between disinformation and the manipulation of information conceived as "malinformation" 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 (Shukla & Pandey, 2025:204). 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 (Buchanan & Imbrie, 2022:199). Ultimately, both unintentional and intentional disinformation, despite not having the same motivation, can have equally harmful effects by distorting reality and confusing people.

One of the main challenges is that the datasets used to train AI often reflect existing prejudices and inequalities in society (Lepage-Richer, 2020:219). 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 (Lindgren, 2023:144). For example, accordingly to Verdegem (2021), 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 (p.303).

In addition to biases in data, the way algorithms are designed and implemented can also contribute to discrimination (Verdegem, 2021). 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 (Lindgren, 2023:142).

Chatbots, defined primarily as “software designed to interact with people through text or voice” (Crowder, 2023:22). Therefore, it is essentially a computer program created to simulate human conversations (Pehcevski, 2024:32). 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 (Crowder, 2023).

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 (Ciesla, 2024:127).

The ease with which this content can be produced and disseminated therefore allows disinformation to spread rapidly, reaching large numbers of people in a short period of time (Rubin, 2022:266). The case of Microsoft's "Tay", which became offensive within 24 hours, demonstrates how chatbots can learn inappropriate behavior from human interactions online (Crowder, 2023).

Crowder (2023) give us another example. It is the case of the chatbot "Tessa" (p. 98), 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 adverse effects and in light of this, XIA emerges as a possibility.

The need for transparent explanations in AI decision-making, or better known as XIA (Bialkova, 2024), 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 (Phillips-Wren et al, 2008:10). This results in a tension between transparency and other objectives such as privacy and intellectual property (Almeida, 2024:172-173).

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 (Bialkova, 2024). 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 (Gaur, 2025:328-329).

Therefore, the lack of transparency and explainability, among others, in the decision-making of Artificial Intelligence (AI) raises problems that undermine the trust and acceptance of this technology. 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 (2024) proposes a Taxonomy to be able to address AI from a responsible approach, combining it with the principles related to XIA or Explained Artificial Intelligence, referring to the following:

Table 1. Principles of Explained Artificial Intelligence (XIA)

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
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, Svetlana, *The rise of AI user applications: chatbots integration foundations and trends*. Springer, 2024, p.190. Available at: https://doi.org/10.1007/978-3-031-56471-0 [Accessed: February 5, 2025].

Accordingly, efforts to create more transparent and explainable AI systems are crucial to foster trust in this technology and ensure 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.

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 (Crowder, 2023). In turn, we have rule-based chatbots that work by following a predefined set of rules and responses (Pehcevski, 2024:60), those based on cognition by simulating human cognitive processes (Crowder, 2023:30) and those based on machine learning and neural networks (Bialkova, 2024:218).

Based on these characteristics, we can propose this typology from the scheme proposed by Bailkova (2024), 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:

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

RELATIONSHIP	APPLICATION OF PRINCIPLE	RISK
EXPLANABILITY		
Uses machine learning algorithms to understand context and learn from past interactions	Minimize that 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		
It 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 understand the context of the question and generate answers.	It depends on the quality of the model and the training data. It can be high if the model is well trained, but it may fail on complex or ambiguous questions.	Over training and data, results can be unpredictable (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		
It 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 and avoid bias, as well as the misuse of captured personal data.	Information asymmetries on complex issues that ultimately depend on the quality of the data.

Source: Prepared according to what was proposed by Bialkova, Svetlana, *The rise of AI user applications: chatbots integration foundations and trends*. Springer, 2024, p.190. Available at: https://doi.org/10.1007/978-3-031-56471-0 [Accessed: February 5, 2025].

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 (Crowder, 2023:8). 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 (Ciesla, 2024:53). In addition, Pehcevski (2024) 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 (p.32).

In 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 (Crowder, 2023:57). Therefore, Replika®'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.

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® is capable of generating 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 (Ciesla, 2024:66-67).

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 (Ciesla, 2024:74). 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 (Ciesla, 2024:77). 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 (Harrington, 2025:19).

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 (XIA) 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 (Ciesla, 2024).	It uses machine learning models, but its internal logic is opaque (Crowder, 2023).	Large language model (LLM) that operates as a "black box", with complex patterns that are difficult to track (Crowder, 2023).	Similar to ChatGPT®, based on complex deep learning models. Its internal logic is difficult to interpret (Ciesla, 2024).	Decision making is complex and reasoning processes are not transparent (Harrington, 2025).
INTERPRETABILITY				
It uses if-then rules, which makes it easy to follow, but the number of rules can make it difficult to maintain (Pehcevski,2024).	Although based on ML, it is oriented towards simpler and more affective interactions than complex information tasks (Crowder, 2023).	Complex, black box-like architecture, difficulty in understanding why it gives a specific response (Crowder, 2023).	Similar to ChatGPT® in its black box nature; opaque internal decision processes (Ciesla, 2024).	Similar to other LLMs, it is characterized by the complexity of its models and the difficulty of tracing its decision processes (Harrington, 2025).

COMPREHENSIBILITY				
It uses flexible pattern matching rules, it may look human, but its understanding is limited (Crowder, 2023).	It tries to be an "online friend," so intelligibility focuses on simulated empathy rather than understanding (Crowder, 2023).	It uses large language models and can generate coherent and contextually relevant text, but can fail at complex reasoning (Crowder, 2023).	Has the ability to search for updated information on the Internet, which improves understanding and response capacity (Ciesla, 2024).	Possibility of errors in coding tasks, problems with using real-world information (Harrington, 2025).
UNDERSTANDABILITY				
If the question is more complex or outside of your range of patterns, you may not respond adequately (Crowder, 2023).	If the user expects deep emotional understanding, they might be disappointed (Crowder, 2023).	A user might ask to write an essay, receiving a well-written text but without understanding whether the information is completely accurate (Crowder, 2023).	Similar to ChatGPT®, it can generate misinformation and responses can be inconsistent (Ciesla, 2024).	It can be opaque in its internal workings, making it difficult to understand and verify (Harrington, 2025).
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 (Ciesla, 2024).	It is not clear how the emotional connections he advertises are formed (Crowder, 2023).	It uses deep learning models, whose internal logic is difficult to interpret (Ciesla, 2024).	Similar to ChatGPT®, based on machine learning models with internal logic that is difficult to understand (Ciesla, 2024).	Shares the risks of ChatGPT® and Bard®: biases, generation of incorrect or false information, and unpredictable behavior (Harrington, 2025).
LEGALITY				
No specific cases of legal problems (Ciesla, 2024).	If it is perceived as a replacement for professional support, there could be legal	It raises issues about the veracity and reliability of information and cannot do causal	No specific legal cases regarding Bard®, but it competes with ChatGPT®, so	No detail of legal cases for DeepSeek®, but being an advanced LLM-

	risks (Crowder, 2023).	analysis or research (Crowder, 2023).	similar risks are assumed (Ciesla, 2024).	based AI technology, it is assumed that it shares the risks of similar models (Harrington, 2025).
ETHICS				
Open source, which allows for community contribution but also means there is less centralized control over its use and potential misuse (Ciesla, 2024).	Relationship with "digital privacy", which points to the specific ethical risk (Ciesla, 2024).	Discussion on deepfakes, which could be associated with ethical risks (Ciesla, 2024).	No specific information on ethical concerns or unique risks, but has parallels with ChatGPT® (Ciesla, 2024).	The fact that it is an open source model implies greater distribution, which could mean greater potential for misuse (Harrington, 2025).

Source: Prepared according to the referenced papers, as well as what was proposed by Bialkova, Svetlana, *The rise of AI user applications: chatbots integration foundations and trends*. Springer, 2024, p.190. Available at: https://doi.org/10.1007/978-3-031-56471-0 [Accessed: February 5, 2025].

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.

3 FINAL CONSIDERATIONS

While AI offers benefits across a number of 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 (XIA) is highlighted to foster trust and ensure

ethical use of AI based on principles such as explainability, interpretability, intelligibility, comprehensibility, transparency, legality and ethics.

XIA is essential for all chatbots, seeking understanding, trust, and accountability. XIA 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 XIA, to mitigate adverse effects and ensure that the technology benefits society.

In conclusion, the ethical and responsible development and implementation of Artificial Intelligence (AI) requires 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 its 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 the technology benefits society as a whole and does not exacerbate existing inequalities or risks. XIA emerges as a fundamental pillar for building a future in which AI is used in an ethical, equitable and responsible manner.

4 REFERENCES

ALKADASH, Tamer, et al., “Maximizing Organizational Efficiency Through HR Information Systems: A Focus on Decision-Making in Tech Firms”. In ALAREENI, Bahaaeddin., & ELGEDAWY, Islam, *AI and business, and innovation research: understanding the potential and risks of AI for modern enterprises*, (1st ed., pp. 431-439), Springer, 2024. Available at: https://doi.org/10.1007/978-3-031-42085-6 [Accessed: February 5, 2025].

ALAREENI, Bahaaeddin., & ELGEDAWY, Islam, *AI and business, and innovation research: understanding the potential and risks of AI for modern enterprises*, (1st ed., pp. 431-439), Springer, 2024. Available at: https://doi.org/10.1007/978-3-031-42085-6 [Accessed: February 5, 2025].

ALMEIDA, I, *Responsible AI in the age of generative models: governance, ethics and risk management* (2024 edition). Now Next Later AI, 2024.

ALSMADI, Izzat, et al., *The NICE cyber security framework: cyber security management*. Springer, 2020. Available at: <https://public.ebookcentral.proquest.com/choice/publicfullrecord.aspx?p=6178526> [Accessed: February 5, 2025].

BHATELE, Kirti Raj, at al., *The emerging role of AI based expert systems in cyber defense and security*, (1st ed., pp. 299-314). Nova Science Publishers, Inc., 2024. Available at: <https://public.ebookcentral.proquest.com/choice/PublicFullRecord.aspx?p=31357405> [Accessed: February 5, 2025].

BIALKOVA, Svetlana, *The rise of AI user applications: chatbots integration foundations and trends*. Springer, 2024. Available at: https://doi.org/10.1007/978-3-031-56471-0 [Accessed: February 5, 2025].

BUCHANAN, Ben, & IMBRIE, Andrew, *The new fire: war, peace, and Democracy in the age of AI*. The MIT Press, 2022. Available at: <https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=2945338> [Accessed: February 5, 2025].

CIESLA, Robert, *The book of chatbots: from ELIZA to ChatGPT*. Springer, 2024. Available at: https://doi.org/10.1007/978-3-031-51004-5 [Accessed: February 5, 2025].

CROWDER, James, *AI chatbots: the good, the bad, and the ugly*. Springer Nature Switzerland, 2024. Available at: https://doi.org/10.1007/978-3-031-45509-4 [Accessed: February 5, 2025].

DAS, Ravindra, *Generative AI Phishing and Cybersecurity Metrics*. Taylor & Francis Group, 2024. Available at: https://public.ebookcentral.proquest.com/choice/PublicFullRecord.aspx?p=31566301 [Accessed: February 5, 2025].

GAUR, Mitisha, "An In-Depth Exploration of Predictive Justice with AI". In POKHARIYAL, Purvi, et al., (Eds.). *AI and emerging technologies: automated decision making, digital forensics and ethical considerations*. (1st ed., pp. 241-272). CRC Press, 2025. Available at: <https://doi.org/10.1201/9781003501152> [Accessed: February 5, 2025].

HALL, Patrick, et al., *Machine learning for high-risk applications: approaches to responsible AI* (First edition), O'Reilly Media, Inc., 2023. Available at: <https://www.oreilly.com/library/view/-/9781098102425/> [Accessed: February 5, 2025].

HARRINGTON, Alexander, THE *DEEPSEEK DISRUPTION: How China's Groundbreaking AI Model Is Transforming the Global Tech Race.* Ria Christie Collections, Uxbridge, UK, 2025.

JHANJHI, Noor, *Utilizing generative AI for cyber defense strategies*. IGI Global, 2025. Available at: https://doi.org/10.4018/979-8-3693-8944-7 [Accessed: February 5, 2025].

LEPAGE-RICHER, Théo, "Adversariality in Machine Learning Systems: On Neural Networks and the Limits of Knowledge". In ROBERGE, Jonathan, & CASTELLE, Michael, *The cultural life of machine learning: an incursion into critical AI studies*. (1st ed., pp. 197-225). Palgrave Macmillan, 2021. Available at: https://doi.org/10.1007/978-3-030-56286-1 [Accessed: February 5, 2025].

LINDGREN, Simon, *Critical theory of AI*. Polity Press, 2024.

Moring, Andreas, *AI on the job: guide to successful human-machine collaboration*. Springer, 2022. Available at: https://doi.org/10.1007/978-3-662-64005-0 [Accessed: February 5, 2025].

MURPHY, Patrick, "A Smart and Secure Healthcare System: Automated Methods for Diagnostics". In BHATELE, Kirti Raj, at al., *The emerging role of AI based expert systems in cyber defense and security*, (1st ed., pp. 299-314). Nova Science Publishers, Inc., 2024. Available at: <https://public.ebookcentral.proquest.com/choice/PublicFullRecord.aspx?p=31357405> [Accessed: February 5, 2025].

PEHCEVSKI, Jovan, *Chatbots and Text generation*. Arcler Press, 2024.

PHILLIPS-WREN, Gloria, et al., *Intelligent decision making: an AI-based approach*. *OhioLINK electronic book center*, Springer, 2008. Available at: <https://doi.org/10.1007/978-3-540-76829-6> [Accessed: February 5, 2025].

POKHARIYAL, Purvi, et al., (Eds.), *AI and emerging technologies: automated decision making, digital forensics and ethical considerations*. (1st ed., pp. 241-272). CRC Press, 2025. Available at: <https://doi.org/10.1201/9781003501152> [Accessed: February 5, 2025].

RADANLIEV, Petar, *The rise of AI Agents: integrating AI, blockchain technologies, and quantum computing,* (First edition), Addison-Wesley, 2025. Available at: <https://www.oreilly.com/library/view/-/9780135352939/> [Accessed: February 5, 2025].

ROBERGE, Jonathan, & CASTELLE, Michael, *The cultural life of machine learning: an incursion into critical AI studies*, (1st ed., pp. 197-225), Palgrave Macmillan, 2021. Available at: <https://doi.org/10.1007/978-3-030-56286-1> [Accessed: February 5, 2025].

RUBIN, Victoria, *Misinformation and Disinformation Detecting Fakes with the Eye and AI*, Springer International Publishing, 2022. Available at: <https://doi.org/10.1007/978-3-030-95656-1> [Accessed: February 5, 2025].

SHUKLA, Divyansh, & PANDEY, Anshul, "An Invisible Threat to the Security of Nations in the Age of "Deepfakes"". In POKHARIYAL, Purvi, et al., (Eds.), *AI and emerging technologies: automated decision making, digital forensics and ethical considerations*. (1st ed., pp. 201-216). CRC Press, 2025. Available at: <https://doi.org/10.1201/9781003501152> [Accessed: February 5, 2025].

VERDEGEM, Pieter, *AI for everyone? critical perspectives*. University of Westminster Press, 2021. Available at: <https://www.jstor.org/stable/10.2307/j.ctv26qjjhj> [Accessed: February 5, 2025].

WRITIUS INC, "The deepseek disruption: How China's groundbreaking AI model is transforming the global tech race", *Writius*, January 31, 2025. Available at: <https://writius.com/the-deepseek-disruption-

how-chinas-groundbreaking-ai-model-is-transforming-the-global-tech-race/> [Accessed: February 5, 2025].