ISBN: 978-65-83849-24-3

RISKS OF DISINFORMATION GENERATED BY AI AND MITIGATION STRATEGIES

https://doi.org/10.63330/aurumpub.021-010

Rodrigo Thomé de Moura[1]

ABSTRACT

This study analyzed the risks of disinformation generated by Artificial Intelligence and presented mitigation strategies capable of reducing its social, political, and institutional impacts. The research aimed to investigate how AI technologies, especially generative models, have expanded the production and circulation of false, misleading, and manipulated content, as well as to assess the consequences of this phenomenon for public trust, democracy, science, and national security. The methodology adopted was bibliographic and qualitative, based on a review of scientific articles, institutional reports, and specialized works discussing Artificial Intelligence, digital disinformation, and informational integrity. The results showed that generative AI enabled the creation of synthetic texts, images, videos, and audio at high speed and scale, making disinformation more sophisticated and harder to detect. It was observed that the increasing realism of deepfakes, voice cloning, and automation through bot networks significantly enhanced the ability to manipulate public perceptions, favoring coordinated campaigns and interference in democratic processes. The analysis also identified that this scenario contributed to the erosion of trust in institutions, the discrediting of scientific evidence, and vulnerabilities in sensitive areas such as public health and national security. The study concluded that mitigating these risks depends on combining technical, political, and educational strategies, including tools for detecting synthetic media, regulations for algorithmic transparency, digital governance policies, and media literacy programs capable of strengthening citizens' critical capacity in the contemporary informational environment.

Keywords: Artificial Intelligence; Disinformation; Deepfakes; Informational security; Mitigation.

[1] Postgraduate in Dance Therapy
Prisma
E-mail: rodrigofsw@gmail.com

INTRODUCTION

The rapid expansion of Artificial Intelligence (AI), especially in its generative applications, has profoundly transformed the contemporary informational landscape, creating new technological possibilities while also triggering significant risks for information circulation and social stability. The advancement of models capable of producing synthetic texts, images, videos, and audio with a high degree of realism has inaugurated a period in which the boundaries between truth and falsification have become increasingly difficult to identify. Authors such as Goodfellow et al. (2016), Wardle and Derakhshan (2017), and Zuboff (2019) have discussed how this phenomenon alters communication dynamics, influences public opinion, and exposes society to new forms of manipulation. In this context, understanding the relationship between AI and disinformation becomes essential to assess its impacts and devise strategies that preserve informational integrity.

This study aimed to analyze the risks of disinformation generated by Artificial Intelligence technologies and discuss mitigation strategies capable of reducing their social, political, and institutional effects. Specifically, it sought to: understand the foundations of generative AI; examine the functioning of digital disinformation and its circulation ecosystem; investigate how AI enhances the creation and dissemination of false content; assess the social and political impacts resulting from this process; and present possible approaches to address the problem. The hypothesis guiding the research considered that AI, despite its benefits, substantially intensifies the reach and realism of disinformation, contributing to the erosion of public trust, democratic manipulation, and vulnerabilities in areas such as health, science, and national security.

The justification for this study lies in the central role that information plays in contemporary societies. In an environment marked by hyperconnectivity, rapid communication flows, and growing dependence on digital platforms, disinformation is not merely a technical problem but a structural phenomenon that threatens democratic processes, public health policies, institutional stability, and social relations. The advancement of generative AI intensifies this scenario, requiring critical and multidisciplinary analyses of its risks and ethical implications.

To develop this investigation, the work was structured into four main parts. After this introduction, the methodology section describes the adopted approach, based on bibliographic research and a qualitative perspective. Next, the development is organized into four subsections: the first presents the fundamental concepts of Generative Artificial Intelligence; the second discusses the different types of digital disinformation and their dynamics on platforms; the third analyzes AI as a vector of disinformation, highlighting scalability, realism, and automation; and the fourth examines the social and political impacts of this phenomenon. Finally, the conclusion revisits the main findings, reaffirms the relevance of the topic, and points to the need for integrated policies and mitigation strategies.

Thus, this research seeks to contribute to the contemporary debate on informational challenges in the age of Artificial Intelligence, offering a broad and critical understanding of the risks involved and reinforcing the importance of ethical, technological, and educational responses that ensure the preservation of truth and social trust.

METHODOLOGY

This research was developed through a qualitative approach and was essentially bibliographic, grounded in the analysis of books, scientific articles, institutional reports, and specialized documents discussing Artificial Intelligence, digital disinformation, and informational security. The choice of this methodology was justified by the conceptual and interpretative nature of the topic, which requires a broad understanding of recent technological transformations and their social and political impacts. Thus, the investigation sought to gather and interpret contributions from authors such as Goodfellow, Wardle, Derakhshan, Zuboff, Floridi, Chesney, and Citron, among other researchers prominent in the contemporary debate on algorithms, synthetic media, and informational integrity.

The methodological path consisted of four main stages. The first involved a systematic review of national and international literature to identify relevant concepts, definitions, and theoretical foundations on generative AI, disinformation ecosystems, and sociopolitical impacts of informational manipulation. Next, case studies, research organization reports, and reference documents produced by entities such as the Council of Europe were selected, addressing phenomena such as deepfakes, coordinated campaigns, and digital interference in democratic processes. The third stage focused on critical analysis of this material, seeking to relate theoretical findings to observed societal transformations, especially regarding the advancement of generative models, the circulation of misleading content, and algorithmic mechanisms that amplify disinformation.

Finally, the results obtained were thematically organized in the development sections, enabling an integrated discussion between technical foundations and social impacts. This methodological structure allowed for a comprehensive understanding of the phenomenon, articulating technological, sociological, and political perspectives, and providing insights for reflection on mitigation strategies that can help address the risks associated with Artificial Intelligence. Thus, the adopted methodology not only theoretically supported the study but also enabled the construction of a consistent, critical analysis aligned with contemporary needs for protecting informational integrity.

DEVELOPMENT

GENERATIVE ARTIFICIAL INTELLIGENCE

Generative Artificial Intelligence has become one of the most innovative and transformative fields in contemporary technology, characterized by its ability to create new and original content based on learning from large volumes of data. Conceptually, it refers to systems capable of generating texts, images, sounds, videos, and other media formats without these contents being previously stored, but rather produced from patterns learned during training. According to Goodfellow et al. (2016), this type of AI relies on probabilistic models capable of understanding complex structures of language and human perception, reproducing them autonomously and in a surprisingly natural way. This feature positions generative AI as a milestone in the history of computing, expanding its use in education, healthcare, design, communication, and entertainment, while also opening fundamental discussions on ethics and responsibility.

Within this category, large-scale language models, known as LLMs (Large Language Models), have become the most popular. They are capable of generating coherent texts, translating content, answering questions, producing summaries, and even simulating specific writing styles. These models, such as GPT, Gemini, and LLaMA, were trained on billions of words, developing the ability to predict the next word in a sequence and thus construct complete narratives. As Kaplan et al. (2020) point out, the larger the volume of data and parameters in a model, the greater its capacity to generate sophisticated responses. Similarly, advances have occurred in image and video generation, with models such as DALL·E, Midjourney, and Stable Diffusion, which can create realistic illustrations, synthetic photographs, and complex scenes based on textual descriptions. These tools have expanded digital creativity, allowing individuals without technical training to produce content previously restricted to highly specialized professionals.

Beyond texts and images, generative AI has also revolutionized audiovisual media through voice synthesis and video manipulation. Voice cloning technologies allow for the replication of any person's tone and timbre with high fidelity, generating entire speeches that appear authentic. Likewise, deepfakes—videos manipulated by deep neural networks—have gained notoriety for their ability to replace faces, synchronize speech, and alter expressions with extreme realism. Studies such as those by Chesney and Citron (2019) highlight that deepfakes represent one of the most concerning forms of synthetic content due to their potential to disseminate disinformation, compromise reputations, manipulate political processes, and create false visual evidence. The sophistication of these methods demonstrates how generative AI can both expand creative tools and generate significant risks for informational integrity and public trust.

In this scenario, it is essential to understand that Generative Artificial Intelligence is not merely a technical advancement but a cultural, social, and ethical phenomenon. As synthetic content production becomes indistinguishable from human production, issues related to authenticity, authorship, privacy, and truth gain centrality in contemporary debate. Thus, studying its basic concepts, models, and applications—including potentially harmful ones—is fundamental to ensuring its use occurs responsibly, transparently, and in alignment with the ethical principles governing digital society.

DIGITAL DISINFORMATION

Digital disinformation has become one of the most urgent problems in contemporary society, intensified by the speed and scale of interactions mediated by digital technologies. In the online environment, disinformation takes different forms, each with specific characteristics and varying degrees of intentionality. So-called fake news are entirely false content deliberately produced to deceive, manipulate, or generate repercussion. There are also manipulated contents, which are based on partially true facts but altered or taken out of context to induce misleading interpretations. Additionally, misleading information—known as misinformation—is shared without explicit intent to cause harm but still contributes to the circulation of false or distorted narratives. Wardle and Derakhshan (2017) classify these phenomena within a spectrum ranging from unintentional error to strategic manipulation, emphasizing that each category requires specific forms of identification and counteraction.

This type of content thrives in a digital ecosystem marked by hyperconnectivity, viral logic, and fragmentation of information sources. On social networks, disinformation circulates rapidly because it finds fertile ground in fast, poorly verified, and emotionally charged interactions. Platforms such as Facebook, X/Twitter, WhatsApp, TikTok, and Instagram function as environments where attractive content has a higher chance of being shared, regardless of its veracity. According to Vosoughi, Roy, and Aral (2018), false news spreads faster than true news precisely because it tends to evoke surprise, indignation, and strong emotional engagement. This means that platform design—structured to maximize attention and engagement—contributes to perpetuating disinformation cycles, often making it difficult for the average user to distinguish between verified information and intentional manipulation.

The role of algorithms and automation in this process is equally central. Recommendation tools based on artificial intelligence select content according to user interest patterns, creating informational bubbles that reinforce pre-existing beliefs and hinder access to diverse perspectives. These algorithms, by privileging engagement, end up amplifying sensationalist or polarizing messages, making disinformation more visible than verifiable content. In parallel, automated systems—such as bots and coordinated accounts—are frequently used to artificially increase the reach of certain content or campaigns. Zannettou et al. (2019) show that organized groups use bot networks to boost false narratives, manipulate public

debates, and influence political processes, turning disinformation into a systemic phenomenon that is difficult to contain.

In this scenario, understanding the complexity of digital disinformation requires analyzing its types, circulation forms, and technological mechanisms that amplify its impact simultaneously. The interaction between users, platforms, and algorithms creates an environment where false or manipulated content achieves high diffusion capacity, with direct effects on public opinion, social trust, and the functioning of democratic institutions. Thus, studying this phenomenon becomes fundamental for developing effective mitigation strategies and public policies that strengthen informational integrity in the digital age.

AI AS A VECTOR OF DISINFORMATION

Artificial Intelligence has established itself as a central vector in the expansion of digital disinformation, mainly due to its ability to produce, replicate, and amplify false content at unprecedented speed and scale. Unlike traditional processes of informational manipulation, which required time, resources, and advanced technical knowledge, AI has made it possible to create falsified texts, images, videos, and audio with just a few commands, democratizing the ability to generate misleading content. Generative tools allow a single individual to produce hundreds or thousands of false messages in minutes, intensifying the volume of circulating disinformation and overloading verification mechanisms. According to Floridi (2021), automation has transformed disinformation into an industrialized phenomenon, capable of spreading rapidly and adapting highly to social media dynamics.

One of the most concerning aspects of this transformation is the growing realism of AI-generated synthetic media. Advanced models produce extremely detailed images, videos that accurately simulate facial expressions, and audio that perfectly imitates the human voice. Deepfakes have become one of the most visible expressions of this capability, enabling the creation of videos in which people appear to say or do things that never happened. As Chesney and Citron (2019) highlight, these contents have the potential to compromise reputations, manipulate political decisions, generate panic in crisis situations, and undermine trust in audiovisual records—a historically central element in validating truth. Although initially detectable by visual flaws, deepfakes have evolved to the point of becoming indistinguishable to the naked eye, significantly increasing their impact.

Beyond direct production of false content, AI amplifies disinformation through automated bots and coordinated campaigns. These systems are programmed to simulate human behavior, participate in debates, boost hashtags, comment on posts, and reinforce specific narratives. Automation allows a small group to create the impression of social consensus or organic engagement around sensitive topics. Zannettou et al. (2019) demonstrate that bot networks have been widely used by political, economic, and

ideological groups to manipulate public perceptions, interfere in electoral debates, and amplify polarizing messages. In many cases, these bots operate in conjunction with generative AI models, creating a highly efficient chain of disinformation production and distribution.

The impacts of this phenomenon are visible in various recent contexts. In electoral processes, AI-fabricated content has been used to defame candidates, manipulate narratives, and influence public opinion, as seen in elections in the United States, India, and several European countries. During health crises, such as the COVID-19 pandemic, synthetic videos and texts were used to spread false information about vaccines, treatments, and preventive measures, contributing to risky behaviors and institutional distrust. In areas such as economics and security, manipulated media have already caused temporary drops in financial markets and spread alarms about nonexistent threats, demonstrating AI's ability to generate concrete and immediate impacts on social life.

Thus, Artificial Intelligence, while representing a significant technological advancement, has also become a powerful catalyst for disinformation. Its ability to produce highly realistic content, operate automatically, and interact with the algorithmic dynamics of digital platforms makes it a central element in understanding contemporary challenges related to informational integrity. Recognizing this role is fundamental for developing mitigation strategies capable of responding to the speed, sophistication, and scale of the problem.

SOCIAL AND POLITICAL IMPACTS

The social and political impacts of disinformation amplified by Artificial Intelligence technologies constitute one of the most serious challenges faced by contemporary societies. The massive circulation of false, manipulated, or misleading content has significantly contributed to the erosion of public trust in institutions, media, and even in the very notion of shared truth. When synthetic information generated by AI becomes indistinguishable from real content, citizens begin to question the credibility of visual evidence, official statements, and verified facts. Zuboff (2019) emphasizes that this scenario of constant uncertainty weakens social cohesion and creates fertile ground for polarizing discourses, making it more difficult to build the minimal consensus necessary for democratic functioning. Thus, trust—the foundation of any structured society—becomes fragile in an environment where the authenticity of information is permanently under suspicion.

AI-supported disinformation also plays a central role in manipulating public opinion and interfering in democratic processes. During electoral periods, falsified content can alter perceptions, shape narratives, and influence the decisions of thousands of voters in short timeframes. Deepfakes, texts fabricated by language models, and coordinated bot networks expand the reach of manipulative campaigns, creating artificial impressions of popular support or spreading false accusations against

political opponents. Wardle and Derakhshan (2017) argue that these mechanisms challenge the fundamental principles of democratic deliberation by distorting public debate, hindering access to reliable information, and strategically manipulating emotions. Digital interference in elections has already been documented in several countries, showing that electoral integrity can be significantly compromised by informational automation technologies.

Beyond the political sphere, the impacts extend to science, public health, and national security. The dissemination of false information about medical treatments, disease diagnosis, or vaccination campaigns—widely observed during the COVID-19 pandemic—has shown how disinformation can generate risky behaviors, reduce adherence to health measures, and aggravate epidemiological crises. In the scientific field, the spread of conspiracy theories and anti-scientific content discredits researchers, hinders evidence-based policies, and reduces the population's ability to differentiate validated knowledge from speculation. In national security, deepfakes and audiovisual manipulations can be used to provoke instability, simulate attacks, create false statements by authorities, or generate collective panic. Chesney and Citron (2019) warn that the sophistication of AI-produced falsifications may render the distinction between real and fabricated threats unfeasible, complicating the rapid response of institutions responsible for social and state protection.

Given this scenario, it becomes evident that AI-enhanced disinformation is not merely a technical problem but a social and political phenomenon of great magnitude. Its effects permeate institutions, decision-making processes, health systems, and social relations, compromising fundamental pillars of democratic coexistence. Understanding these impacts is an essential step toward formulating effective policies for mitigation, strengthening the population's digital literacy, and ensuring that emerging technologies are used ethically and responsibly..

CONCLUSION

This study made it possible to understand that Artificial Intelligence, especially in its generative applications, has played a decisive role in transforming the contemporary informational ecosystem. The ability of these technologies to produce synthetic texts, images, videos, and audio with a high degree of realism has profoundly altered the dynamics of information circulation, expanding the potential for creating and disseminating false, manipulated, or misleading content. The analysis revealed that AI not only accelerated these processes but also made them more complex, efficient, and difficult to detect, shaping a disinformation scenario more powerful and sophisticated than any previously observed.

The results indicated that disinformation generated or amplified by AI directly contributed to the erosion of public trust—a phenomenon that affects democratic institutions, communication systems, scientific practices, and social relations. The growing presence of deepfakes, synthetic speeches,

manipulated images, and automated campaigns has made it harder to distinguish between true and false content, increasing the sense of uncertainty and informational vulnerability. Furthermore, it was highlighted that these technologies facilitated the manipulation of public opinion and political interference, creating conditions for coordinated campaigns that significantly affect electoral processes, governmental agendas, and public debates. It was also evident that AI-supported disinformation caused harm to science, health, and national security, especially in crisis contexts such as pandemics and geopolitical instabilities.

Given this scenario, the study demonstrated that the risks associated with AI do not reside solely in its technical capacity but in the way these technologies are incorporated into social and political dynamics. Thus, mitigating these risks depends on integrated and multidimensional actions. The strategies analyzed indicated the need to combine technical solutions—such as deepfake detection tools, synthetic content labeling, and authentication systems—with regulatory policies that establish algorithmic transparency, platform accountability, and governance mechanisms that protect informational integrity. Likewise, the importance of media education was emphasized as a fundamental strategy to strengthen citizens' critical capacity, enabling them to recognize signs of manipulation and develop greater autonomy when faced with misleading content.

It is concluded, therefore, that addressing AI-generated disinformation requires a joint effort among governments, researchers, technology companies, educational institutions, and civil society. Only through such cooperation will it be possible to ensure that Artificial Intelligence is employed ethically, responsibly, and in alignment with democratic values, preserving social trust and informational security in an increasingly complex and challenging digital context..

REFERENCES

1. Chesney, Robert; Citron, Danielle. Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, v. 107, p. 1753–1819, 2019. Available at: https://www.californialawreview.org/print/deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security/. Accessed on: 14 Nov. 2025.

2. Floridi, Luciano. The Ethics of Artificial Intelligence. Oxford: Oxford University Press, 2021.

3. Goodfellow, Ian; Bengio, Yoshua; Courville, Aaron. Deep Learning. Cambridge: MIT Press, 2016.

4. Kaplan, Jared et al. Scaling laws for neural language models. arXiv, 2020. Available at: https://arxiv.org/abs/2001.08361. Accessed on: 14 Nov. 2025.

5. Vosoughi, Soroush; Roy, Deb; Aral, Sinan. The spread of true and false news online. Science, v. 359, n. 6380, p. 1146–1151, 2018. Available at: https://www.science.org/doi/10.1126/science.aap9559. Accessed on: 14 Nov. 2025.

6. Wardle, Claire; Derakhshan, Hossein. Information Disorder: Toward an Interdisciplinary Framework. Strasbourg: Council of Europe, 2017. Available at: https://firstdraftnews.org/wp-content/uploads/2017/11/PREMS-162317-GBR-2018-Report-de%CC%81sinformation-1.pdf. Accessed on: 14 Nov. 2025.

7. Zannettou, Savvas et al. Disinformation warfare: Understanding state-sponsored trolls on Twitter and their influence on the web. In: Companion Proceedings of the World Wide Web Conference (WWW), 2019. Available at: https://arxiv.org/pdf/1801.09288. Accessed on: 14 Nov. 2025.

8. Zuboff, Shoshana. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. New York: PublicAffairs, 2019.